

Forecast Reconciliation with Uncertainty Quantification

Jan Kloppenborg Møller^{a,*}, Peter Nystrup^a, Poul G. Hjorth^a, Henrik Madsen^a

^a*Department of Applied Mathematics and Computer Science, Technical University of Denmark, Denmark*

Abstract

We propose a new method for estimating the weight matrix used for forecast reconciliation as parameters in a general linear model. This enables quantification of the weight matrix uncertainty, which is important because it propagates to the variance of the reconciled forecast. More specifically, we show that forecast reconciliation can be formulated as an orthogonal projection from the space of base-forecast errors into a coherent linear subspace. We use variance decomposition to derive the central estimator for the forecast-error covariance matrix and prove that distance-reducing properties apply to the reconciled forecasts at all levels of the hierarchy as well as to the forecast-error covariance. We show how shrinkage can be introduced by imposing specific priors on the weight matrix and the forecast-error covariance matrix. The method is illustrated in a simulation study that shows consistent improvements in the log-score. Finally, standard errors for the weight matrix and the variance-separation formula are illustrated using a case study of forecasting electricity load in Sweden.

Keywords: Forecast reconciliation, general linear model, orthogonal projection, restricted maximum likelihood, maximum a posteriori estimation, shrinkage estimation

1. Introduction

Forecast reconciliation is an effective way to ensure coherency across hierarchies defined by linear constraints. Forecasts for a hierarchy are coherent when they fulfil its constraints. Classical examples include that aggregated sales forecasts for individual stores must align with regional and

*Corresponding author

Email address: jkmo@dtu.dk (Jan Kloppenborg Møller)

national sales forecasts (structural reconciliation, see, e.g., Athanasopoulos et al., 2009) and quarterly forecasts must align with biannual and annual forecasts (temporal reconciliation, see, e.g., Athanasopoulos et al., 2017). In addition to ensuring coherency, reconciliation often improves forecast accuracy on all levels of a hierarchy.

Pritularga et al. (2021) argued that the effect of uncertainty in forecast reconciliation has been overlooked. By decomposing the variance of the reconciled forecasts into different sources, they showed that uncertainties propagate from the variance–covariance matrix estimation to the reconciliation weights, thereby increasing the uncertainty of the reconciled forecasts. Multiple studies have found that more complete approximations of the variance–covariance matrix improve the accuracy of the reconciled forecasts at the cost of increased variance (Nystrup et al., 2021; Panagiotelis et al., 2021; Pritularga et al., 2021; Møller et al., 2024). To the best of our knowledge, we are the first to quantify the uncertainty of the estimated forecast reconciliation weights and apply this to obtain better estimates of the forecast-error covariance.

Since its introduction by Hyndman et al. (2011), several articles have improved the theoretical understanding of forecast reconciliation. Wickramasuriya et al. (2019) showed the relation between generalized least squares and the minimum trace (minT) solution and derived the variance of the reconciled forecast errors. Panagiotelis et al. (2021) presented a geometric interpretation of forecast reconciliation and proved that the minT solution is optimal in the sense that it minimises the expected loss. Di Fonzo and Girolimetto (2024) extended the work by Hollyman et al. (2021) on the connection between forecast combination and reconciliation to include linear constraints.

Wickramasuriya et al. (2019) provided theoretical justification for using the variance–covariance matrix of the base forecast errors as an estimate of the unknown and unidentifiable variance–covariance matrix for the coherency errors. However, the dimension of this matrix is in general large and it is often ill-conditioned. Therefore, shrinkage or simplification (Athanasopoulos et al., 2017) is usually applied. Nystrup et al. (2021) used eigenvalue decomposition and Møller et al. (2024) proposed a parameterised likelihood approach to reduce the dimension of the problem. Eckert et al. (2021) used Bayesian methods to find the posterior distribution of the reconciled forecasts and bias shrinkage to shrink some weights towards zero. By now, the minT approach, including optimal shrinkage (Ledoit and Wolf, 2003), is established as the standard for forecast reconciliation.

This article introduces a number of new results and insights on the statistical properties of forecast reconciliation: 1) we show how the reconciliation weights can be estimated as parameters in a general linear model (LM) independently of the forecast variance; 2) the LM formulation implies an orthogonal projection that is shown to generalise to all levels of the hierarchy through the coherency constraints; 3) from the orthogonal projection and the resulting analysis-of-variance (ANOVA) type separation of variation, we derive a central estimator for the forecast error variance–covariance matrix; 4) we prove that distance-reducing properties apply to the reconciled forecasts at all levels of the hierarchy as well as to the forecast-error covariance; 5) we formulate forecast reconciliation with shrinkage using maximum a posteriori (MAP) estimation, which highlights the choice of priors; and 6) the LM formulation allows us to estimate the parameter (weight matrix) variance–covariance matrix and obtain a better estimate of the forecast-error variance–covariance. The introduced formulations as well as the analysis of the statistical properties of forecast reconciliation pave the way for future work on statistical testing and modelling of the parameters in the reconciliation weight matrix.

The outline of this article is as follows. In Section 2, we formulate the general framework, show its equivalence with forecast reconciliation, and derive a number of in-sample results related to orthogonal projection. We show how shrinkage can be formulated as a MAP estimation problem and derive the mean-value parameters and forecast variances in Section 3. Results on parameter and forecast variance are presented in Section 4 along with some test statistics. Section 5 analyses improvements in forecast distributions in a simulation study. Section 6 illustrates the variance separation and standard errors for the high-dimensional weight matrix in a case study on electricity load forecasting in Sweden. Finally, Section 7 concludes.

A full list of symbols used in this article is given in Appendix A. As many results in this work involve Kronecker products and vectorisation, we have listed the most important relations in Appendix B. Multivariate distributions and their usage in objective functions are briefly presented in B.2.

2. Forecast reconciliation as a linear model

Hyndman et al. (2011) proposed to formulate the forecast reconciliation problem as

$$\hat{\mathbf{y}}_t = \mathbf{S}\tilde{\mathbf{y}}_t + \boldsymbol{\epsilon}_t^h; \quad \boldsymbol{\epsilon}_t^h \sim N(\mathbf{0}, \boldsymbol{\Sigma}_h), \quad (1)$$

where $\hat{\mathbf{y}}_t \in \mathbb{R}^n$ is the base forecast on all levels at time t , $\mathbf{S} \in \mathbb{R}^{n \times m}$ is a summation matrix ensuring coherency, $\tilde{\mathbf{y}}_t \in \mathbb{R}^m$ is the reconciled forecast, $\epsilon_t^h \in \mathbb{R}^n$ is the reconciliation error, $\Sigma_h \in \mathbb{R}^{n \times n}$ is the variance–covariance of the reconciliation error, n is the dimension of the hierarchy, and m is the dimension of the bottom level of the hierarchy.

As a simple example, consider forecasts of quarterly, biannual, and annual observations $\hat{\mathbf{y}}_t = [\hat{y}_{A,t}, \hat{y}_{H_1,t}, \hat{y}_{H_2,t}, \hat{y}_{Q_1,t}, \dots, \hat{y}_{Q_4,t}]^\top$. In this case, $m = 4$, $n = 7$, and the summation matrix is

$$\mathbf{S} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ & & & \mathbf{I}_4 \end{bmatrix} = \begin{bmatrix} \mathbf{S}_{\text{top}} \\ \mathbf{I}_4 \end{bmatrix}. \quad (2)$$

We refer to $\mathbf{S}_{\text{top}}$ as the top-level summation matrix.

If Σ_h were known, the formulation (1) would lead to estimated reconciled forecasts

$$\tilde{\mathbf{y}}_t = (\mathbf{S}^\top \Sigma_h^{-1} \mathbf{S})^{-1} \mathbf{S}^\top \Sigma_h^{-1} \hat{\mathbf{y}}_t = \mathbf{P} \hat{\mathbf{y}}_t. \quad (3)$$

The variance–covariance matrix Σ_h is unknown in real applications and it is unidentifiable (Wickramasuriya et al., 2019). Usually, some version of the observed base-forecast-error variance–covariance matrix is used a proxy.

As pointed out by Panagiotelis et al. (2021), the matrix $\mathbf{S}\mathbf{P}$ is a projection matrix from the n -dimensional base forecast ($\hat{\mathbf{y}}_t \in \mathbb{R}^n$) into a coherent m -dimensional linear subspace. In the following, we will view the weight matrix, $\mathbf{P}$, as parameters in a general LM. It follows directly from the construction of $\mathbf{P}$ that there is a set of linear constraints built into the estimation (Wickramasuriya et al., 2019)

$$\mathbf{P}\mathbf{S} = \begin{bmatrix} \mathbf{P}_{\text{top}} & \mathbf{P}_{\text{bot}} \end{bmatrix} \begin{bmatrix} \mathbf{S}_{\text{top}} \\ \mathbf{I} \end{bmatrix} = \mathbf{I}_m, \quad (4)$$

where $\mathbf{P}_{\text{top}} \in \mathbb{R}^{m \times (n-m)}$ and $\mathbf{P}_{\text{bot}} \in \mathbb{R}^{m \times m}$ are the weight matrices for the top and bottom level, respectively. Equation (4) gives m^2 linear constraints, and with $\mathbf{P} \in \mathbb{R}^{m \times n}$ we have a total of $nm - m^2 = m(n - m)$ free parameters to be estimated. Below, we will show how this can be formulated as a linear regression problem with linear constraints, thereby allowing estimation of the parameters and reconciled variance at the bottom level using maximum likelihood (ML) or restricted maximum likelihood (REML) estimation.

2.1. Maximum likelihood and restricted maximum likelihood

The main contribution of this article is to show that the forecast reconciliation problem can be written as a LM of the bottom-level base-forecast errors

$$\mathbf{y} - \hat{\mathbf{y}}_{\text{bot}} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}; \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma}), \quad (5)$$

where $\mathbf{y} \in \mathbb{R}^{mT}$ is the collection of all bottom-level observations, $\hat{\mathbf{y}}_{\text{bot}} \in \mathbb{R}^{mT}$ is the collection of all bottom-level base forecasts, T is the number of observations in the training set, $\boldsymbol{\epsilon}$ is the residuals from the LM, and the variance–covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{mT \times mT}$ and design matrix $\mathbf{X} \in \mathbb{R}^{mT \times p}$ are chosen in appropriate ways.

Specifically, we show in Theorem 2.2 (in Section 2.2) that the reconciliation problem can be formulated as the LM

$$\mathbf{y}_t - \hat{\mathbf{y}}_{\text{bot},t} = \left[\mathbf{I}_m \otimes (\hat{\mathbf{y}}_{\text{top},t}^\top - \hat{\mathbf{y}}_{\text{bot},t}^\top \mathbf{S}_{\text{top}}^\top) \right] \boldsymbol{\beta}_{\text{top}} + \boldsymbol{\epsilon}_t; \quad \boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma}_r), \quad (6)$$

for a specific time point t , with $\mathbf{y}_t \in \mathbb{R}^m$, $\hat{\mathbf{y}}_{\text{bot},t} \in \mathbb{R}^m$, $\hat{\mathbf{y}}_{\text{top},t} \in \mathbb{R}^{n-m}$, $\boldsymbol{\beta}_{\text{top}} \in \mathbb{R}^{m(n-m)}$, $\boldsymbol{\epsilon}_t \in \mathbb{R}^m$, and $\boldsymbol{\Sigma}_r \in \mathbb{R}^{m \times m}$. Section 2.2 provides some general results needed to prove the result given in (6). In Section 2.3, we show that if $\boldsymbol{\Sigma}_h$ were chosen as the observed base-forecast-error variance–covariance, then the ML estimate for $\boldsymbol{\beta}_{\text{top}}$ is equivalent to the weight matrix $\mathbf{P}$ in (3).

Note that (1) is a model for the coherency error, while (6) focuses on the bottom-level forecast errors and, as such, the observed values are included on the left-hand side of (5). Unlike the variance–covariance of the reconciliation error $\boldsymbol{\Sigma}_h$, the variance–covariance of the regression error $\boldsymbol{\Sigma}_r$ can be identified from the observed data (provided that sufficient data is available). The formulation as a LM allows us to prove a number of in-sample distance-reducing properties in Section 2.4. Later, in Section 3, we will show how shrinkage can be introduced by imposing specific priors on $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_r$ using MAP estimation.

2.2. Results for a class of general linear models

First, we consider a class of LMs that is more general than the specific models needed to show the results related to forecast reconciliation. We begin by considering one particular time of prediction and formulate the following linear regression model

$$\mathbf{y}_t = \mathbf{X}_{\cdot,t} \boldsymbol{\beta} + \boldsymbol{\epsilon}_t; \quad \boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma}_r) \text{ and iid.}, \quad (7)$$

where $\mathbf{X}_{\cdot,t} \in \mathbb{R}^{m \times p}$ is the rows of the design matrix $\mathbf{X} \in \mathbb{R}^{nT \times p}$ corresponding to the observations $\mathbf{y}_t$, and $\boldsymbol{\epsilon}_t \in \mathbb{R}^m$ is a sequence of iid. normal random vectors with mean zero and variance-covariance matrix $\boldsymbol{\Sigma}_r$.

Reconciled forecasts for other levels are constructed by aggregating the forecasts for the bottom level ($\tilde{\mathbf{y}}_t = \mathbf{X}_{\cdot,t} \hat{\boldsymbol{\beta}}$), i.e.,

$$\tilde{\mathbf{y}}_{\text{full},t} = \mathbf{S} \tilde{\mathbf{y}}_t = \mathbf{S} \mathbf{X}_{\cdot,t} \hat{\boldsymbol{\beta}}, \quad (8)$$

where $\tilde{\mathbf{y}}_{\text{full},t} \in \mathbb{R}^n$ is the collection of reconciled forecasts on all top and bottom levels. This implies that the forecasts are coherent for all choices of $\mathbf{X}_{\cdot,t}$.

With $\mathbf{y}$ as defined in (5) and

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_{\cdot,T} \\ \mathbf{X}_{\cdot,T-1} \\ \vdots \\ \mathbf{X}_{\cdot,1} \end{bmatrix}, \quad (9)$$

the full model for all bottom-level observations, assuming independence between the T (m -dimensional) observations, is

$$\mathbf{y} = \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\epsilon}; \quad \boldsymbol{\epsilon} \sim N(0, \mathbf{I}_T \otimes \boldsymbol{\Sigma}_r). \quad (10)$$

We consider design matrices of the following special form

$$\mathbf{X}_{\cdot,t} = \begin{bmatrix} \mathbf{x}_{1,t}^\top & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_{2,t}^\top & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{x}_{m,t}^\top \end{bmatrix}, \quad (11)$$

with $\mathbf{x}_{i,t} \in \mathbb{R}^{p_i}$, meaning that, in general, we allow $\mathbf{x}_{i,t} \neq \mathbf{x}_{j,t}$. This will allow us to remove columns of the design matrix as part of a test strategy. We will sometimes need the design matrix $\mathbf{X}_{i,\cdot}$ corresponding to a specific bottom-level observation

$$\mathbf{X}_{i,\cdot} = \begin{bmatrix} \mathbf{x}_{i,T}^\top \\ \mathbf{x}_{i,T-1}^\top \\ \vdots \\ \mathbf{x}_{i,1}^\top \end{bmatrix}. \quad (12)$$

The general solution to the model defined by (9) is well-know, and the ML or generalized least square (GLS) estimate for the mean-value parameter estimates $\hat{\beta}$ in model (10) is

$$\hat{\beta} = (\mathbf{X}^\top (\mathbf{I}_T \otimes \hat{\Sigma}_r^{-1}) \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{I}_T \otimes \hat{\Sigma}_r^{-1}) \mathbf{y}. \quad (13)$$

Further, with $\mathbf{e}_t = \mathbf{y}_t - \mathbf{X}_{\cdot,t} \hat{\beta}$, the ML estimate of Σ_r , for a given $\hat{\beta}$, is

$$\hat{\Sigma}_{r,\text{ML}} = \frac{1}{T} \sum_{t=1}^T \mathbf{e}_t \mathbf{e}_t^\top. \quad (14)$$

Hence, in general the mean-value parameters $\hat{\beta}$ depend on the estimate of Σ_r , which depends on the estimate of β . In situations similar to the one usually considered in forecast reconciliation, the solution simplifies considerably with the precise result given in the theorem below.

Theorem 2.1 (LM/GLS solution). In the special case where, for all (i, j) , $\mathbf{x}_{i,t} = \mathbf{x}_{j,t} \in \mathbb{R}^{\bar{p}}$, the parameter estimates

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}, \quad (15)$$

are independent of Σ_r and the REML estimate of Σ_r is

$$\hat{\Sigma}_{r,\text{REML}} = \frac{1}{T - \bar{p}} \sum_{t=1}^T \mathbf{e}_t \mathbf{e}_t^\top, \quad (16)$$

where $\mathbf{e}_t = \mathbf{y}_t - \mathbf{X}_{\cdot,t} \hat{\beta}$.

Proof. See Appendix C. ■

The main takeaway from Theorem 2.1 is that the parameter estimates (15), which will be the forecast reconciliation weights, are independent of the choice of Σ_r .

Note that in the case where $\mathbf{x}_{i,t} \neq \mathbf{x}_{j,t}$ for some (i, j) , then the independence between the mean-value and variance parameters does not hold, and the REML estimate is given as the solution to

$$\hat{\Sigma}_{r,\text{REML}} = \frac{1}{T} \left(\sum_{t=1}^T \mathbf{e}_t \mathbf{e}_t^\top + \frac{\partial \log |\mathbf{X}^\top (\mathbf{I}_T \otimes \Sigma_r^{-1}) \mathbf{X}|}{\partial \Sigma_r^{-1}} \bigg|_{\Sigma_r^{-1} = \hat{\Sigma}_{r,\text{REML}}^{-1}} \right). \quad (17)$$

The details for arriving at (17) are given in Appendix C. In the general case, where the parameters $\hat{\beta}$ depend on the estimate of Σ_r , the relaxation algorithm (see e.g. Madsen, 2008) can be used to iterate between (13) and (14) or (17) until convergence.

In the considered setting, both $\mathbf{X}$ and Σ are high-dimensional sparse matrices, but the products can be formulated as lower-dimensional dense matrices (see Appendix C). In the special case of $\mathbf{x}_{i,t} = \mathbf{x}_{j,t}$, the solution is given directly by (15) and (16) (or the ML version of the latter).

It follows directly from (the proof of) Theorem 2.1 that in the special case where $\forall (i, j) \mathbf{x}_{i,t} = \mathbf{x}_{j,t} = \mathbf{x}_{j,t} \in \mathbb{R}^{\bar{p}}$, implying that $\mathbf{X}_{i,\cdot} = \mathbf{X}_{j,\cdot} = \mathbf{X}_{1,\cdot}, \forall (i, j)$, then the estimate of β can be formulated as

$$\hat{\beta}^m = (\mathbf{X}_{1,\cdot}^\top \mathbf{X}_{1,\cdot})^{-1} \mathbf{X}_{1,\cdot}^\top \mathbf{Y}, \quad (18)$$

where $\mathbf{Y} = \text{vec}^{-1}(\mathbf{y}) \in \mathbb{R}^{T \times m}$ and $\hat{\beta}^m = [\hat{\beta}_1 \dots \hat{\beta}_m]^\top \in \mathbb{R}^{m \times \bar{p}}$ does not depend on Σ_r . The result in (15) is given in terms products of sparse matrices (multiplied by a vector), while (18) is given in terms of products of lower-dimensional dense matrices multiplied by a matrix.

2.3. Equivalence to forecast reconciliation

The weight matrix $\mathbf{P}$ depends on the choice of proxy for the variance–covariance matrix Σ_h . Consequently, the equivalence between the regression model (5) and forecast reconciliation (1) and (3) depends on that choice. The next theorem states the equivalence when Σ_h is chosen as the observed base-forecast error variance–covariance matrix.

Theorem 2.2 (Equivalence to forecast reconciliation). The model (5), with $\mathbf{X}_{\cdot,t} = \mathbf{I} \otimes \hat{\mathbf{y}}_t^\top$ and linear coherency constraints $\text{vec}^{-1}(\beta)^\top \mathbf{S} = \mathbf{I}_m$, is equivalent to forecast reconciliation. Formally

$$\mathbf{P} = \text{vec}^{-1}(\hat{\beta})^\top, \quad (19)$$

where $\mathbf{P}$ is the usual forecast reconciliation weight matrix (3) using

$$\hat{\Sigma}_h = \frac{1}{T} (\mathbf{Y} \mathbf{S}^\top - \hat{\mathbf{Y}})^\top (\mathbf{Y} \mathbf{S}^\top - \hat{\mathbf{Y}}) = \frac{1}{T} \sum_{i=1}^T (\mathbf{S} \mathbf{y}_t - \hat{\mathbf{y}}_t)(\mathbf{S} \mathbf{y}_t - \hat{\mathbf{y}}_t)^\top. \quad (20)$$

where $\mathbf{Y} = \text{vec}^{-1}(\mathbf{y})$, and $\hat{\mathbf{Y}} = [\hat{\mathbf{Y}}_{\text{top}} \quad \hat{\mathbf{Y}}_{\text{bot}}] = [\text{vec}^{-1}(\hat{\mathbf{y}}_{\text{top}}) \quad \text{vec}^{-1}(\hat{\mathbf{y}}_{\text{bot}})]$. The model, including coherency constraints, can be formulated as

$$\mathbf{y}_t - \hat{\mathbf{y}}_{\text{bot},t} = \left[\mathbf{I}_m \otimes (\hat{\mathbf{y}}_{\text{top},t}^\top - \hat{\mathbf{y}}_{\text{bot},t}^\top \mathbf{S}_{\text{top}}^\top) \right] \beta_{\text{top}} + \epsilon_t; \quad \epsilon_t \sim N(\mathbf{0}, \Sigma_r), \quad (21)$$

with $\mathbf{S} = [\mathbf{S}_{\text{top}}^\top \quad \mathbf{I}]^\top$.

Further, with

$$\mathbf{e}_t = \mathbf{y}_t - \hat{\mathbf{y}}_{\text{bot},t} - \left[\mathbf{I}_m \otimes (\hat{\mathbf{y}}_{\text{top},t}^\top - \hat{\mathbf{y}}_{\text{bot},t}^\top \mathbf{S}_{\text{top}}^\top) \right] \hat{\boldsymbol{\beta}}_{\text{top}} = \mathbf{y}_t - \tilde{\mathbf{y}}_t, \quad (22)$$

the ML and REML estimates of $\boldsymbol{\Sigma}_r$ are

$$\hat{\boldsymbol{\Sigma}}_{r,\text{ML}} = \frac{1}{T} \sum_{i=1}^T \mathbf{e}_i \mathbf{e}_i^\top; \quad \hat{\boldsymbol{\Sigma}}_{r,\text{REML}} = \frac{1}{T - (n - m)} \sum_{i=1}^T \mathbf{e}_i \mathbf{e}_i^\top, \quad (23)$$

and the estimate $\hat{\boldsymbol{\beta}}_{\text{top}}$ is independent of $\boldsymbol{\Sigma}_r$.

Proof. See Appendix D. ■

Different proxies for $\boldsymbol{\Sigma}_h$ can be used, but in this section we will use the observed variance–covariance of the base-forecast errors (20). Equation (21) is formulated in terms of the LM given in (5) or (7) with $\mathbf{X}_{:,t} = \mathbf{I}_m \otimes (\hat{\mathbf{y}}_{\text{top},t}^\top - \hat{\mathbf{y}}_{\text{bot},t}^\top \mathbf{S}_{\text{top}}^\top)$, where the full design matrix $\mathbf{X}$ is constructed by (9). The model (21) defines the reconciliation problem as a linear regression problem and, consequently, as a linear projection of the bottom-level base-forecast errors. The dimension of the model, which is the dimension of $\boldsymbol{\beta}_{\text{top}}$, is $m \cdot (n - m)$.

The residual error $\boldsymbol{\epsilon}_t$ is (in the case of a perfectly specified model) referred to as the irreducible error by Pritularga et al. (2021), while variance components related to estimation of the base-forecast variance–covariance matrix appear as the variance of the estimator $\hat{\boldsymbol{\beta}}_{\text{top}}$. A bias correction can be included in (21) as $\mathbf{X}_{:,t} = \mathbf{I}_m \otimes [1, \hat{\mathbf{y}}_{\text{top},t}^\top - \hat{\mathbf{y}}_{\text{bot},t}^\top \mathbf{S}_{\text{top}}^\top]$ at the expense of m extra parameters. Wickramasuriya et al. (2019) showed that if $\boldsymbol{\Sigma}_h$ were known, then the variance of the (bottom-level) reconciled forecasts would be given by

$$V[\mathbf{y}_t - \tilde{\mathbf{y}}_t] = \mathbf{P} \boldsymbol{\Sigma}_h \mathbf{P}^\top. \quad (24)$$

The following corollary shows the equivalence between (24) and the ML estimate of $\boldsymbol{\Sigma}_r$ in (23), which is biased downwards compared with the REML estimator (23), which is central.

Corollary 2.1 (Reconciled variance). If $\boldsymbol{\Sigma}_h$ is chosen as $\hat{\boldsymbol{\Sigma}}_h = \frac{1}{T} (\mathbf{Y} \mathbf{S}^\top - \hat{\mathbf{Y}})^\top (\mathbf{Y} \mathbf{S}^\top - \hat{\mathbf{Y}})$, then the ML estimate of the residual variance in model (21) is equal to the variance of the reconciled forecasts (24):

$$\mathbf{P} \hat{\boldsymbol{\Sigma}}_h \mathbf{P}^\top = \hat{\boldsymbol{\Sigma}}_{r,\text{ML}}. \quad (25)$$

Proof. See Appendix E. ■

Theorem 2.2 and Corollary 2.1 show that the forecast reconciliation problem using the full forecast error variance–covariance matrix as proxy for Σ_h can be formulated as a LM for the bottom-level base-forecast errors. One of the key benefits is that we get access to all the well-know statistical methods associated with the general linear model, such as the in-sample distance-reducing properties presented in the next section.

2.4. Distance-reducing properties

A key component in LMs is the separation of variation and the resulting (ANOVA-type) tests for model reduction that are based in orthogonal projections. In this section we show how reconciliation can be formulated as orthogonal projections and prove distance-reducing properties. In Theorem 2.3, we show that

$$\|\mathbf{y}_I - \tilde{\mathbf{y}}_I\|^2 \leq \|\mathbf{y}_I - \hat{\mathbf{y}}_I\|^2, \quad (26)$$

where I corresponds to the collection of all observations/forecasts at a particular level.

When $\Sigma = \sigma^2 \mathbf{I}$, the LM is often formulated in terms of orthogonal projections (see e.g. Madsen and Thyregod, 2011). For more general variance–covariance structures, the LM defines orthogonal projections from a transformed space. The precise definitions are well-known from statistical theory, but given in the remark below for completeness.

Remark 2.1 (LM projection). With $\mathbf{z} = \mathbf{y} - \hat{\mathbf{y}}_{\text{bot}} \in \mathcal{Y} \subseteq \mathbb{R}^{mT}$, the estimate $(\hat{\beta})$ based on the model (5) defines an orthogonal projection from the transformed space $\mathcal{Y}^*$, with $\mathbf{z}^* \in \mathcal{Y}^* \subseteq \mathbb{R}^{mT}$ and $\mathbf{z}^* = \Sigma^{-1/2} \mathbf{z}$, $\Sigma = \Sigma^{1/2} \left(\Sigma^{1/2} \right)^\top$, into a p -dimensional linear subspace. If $\Sigma = \sigma^2 \mathbf{I}$, then the estimate $\hat{\beta}$ defines an orthogonal projection from $\mathcal{Y}$ into a p -dimensional linear subspace.

Using Remark 2.1, the model (10)–(11) defines an orthogonal projection from $\mathbb{R}^{Tm}$ into a $\bar{p}m$ dimensional linear subspace. The next lemma states the LM as a projection of the base-forecast errors and shows that, due to the coherency constraints, the projection matrix is the same for base-forecast errors at all levels of the hierarchy.

Lemma 2.1 (Projection and separation of variation). Assume that $\mathbf{X}_{\cdot,t}$ is chosen as in Theorem 2.2 and that $\text{rank}(\mathbf{X}_{1,\cdot}) = n - m$. If $\mathbf{H} = \mathbf{X}_{1,\cdot}(\mathbf{X}_{1,\cdot}^\top \mathbf{X}_{1,\cdot})^{-1} \mathbf{X}_{1,\cdot}^\top$, then $\mathbf{H}$ defines an orthogonal projection for any selection of columns $I \subseteq \{1, 2, \dots, n\}$ of $\hat{\mathbf{Y}} = \text{vec}^{-1}(\hat{\mathbf{y}}) \in \mathbb{R}^{T \times n}$ and corresponding selection of rows of $\mathbf{S}$. In particular, using the notation $\hat{\mathbf{Y}}_I := \hat{\mathbf{Y}}_{\cdot,I}$ and $\mathbf{S}_I := \mathbf{S}_{I,\cdot}$, then $\mathbf{X}_{1,\cdot}$ defines an orthogonal projection and

$$\begin{aligned} (\mathbf{Y} \mathbf{S}_I^\top - \hat{\mathbf{Y}}_I)^\top (\mathbf{Y} \mathbf{S}_I^\top - \hat{\mathbf{Y}}_I) &= \mathbf{S}_I (\mathbf{Y} - \tilde{\mathbf{Y}})^\top (\mathbf{Y} - \tilde{\mathbf{Y}}) \mathbf{S}_I^\top + \\ &\quad (\tilde{\mathbf{Y}} \mathbf{S}^\top - \hat{\mathbf{Y}}_I)^\top (\tilde{\mathbf{Y}} \mathbf{S}^\top - \hat{\mathbf{Y}}_I) \end{aligned} \quad (27a)$$

$$\begin{aligned} &= (\mathbf{Y} \mathbf{S}_I^\top - \hat{\mathbf{Y}}_I)^\top (\mathbf{I} - \mathbf{H}) (\mathbf{Y} \mathbf{S}_I^\top - \hat{\mathbf{Y}}_I) + \\ &\quad (\mathbf{Y} \mathbf{S}_I^\top - \hat{\mathbf{Y}}_I)^\top \mathbf{H} (\mathbf{Y} \mathbf{S}_I^\top - \hat{\mathbf{Y}}_I), \end{aligned} \quad (27b)$$

where $\tilde{\mathbf{Y}} = \text{vec}^{-1}(\tilde{\mathbf{y}}) \in \mathbb{R}^{T \times m}$ is the collection of all bottom-level reconciled forecasts.

Proof. See Appendix F. ■

The sum-of-squares decomposition (27a) can be written in terms of the variance-covariance matrices discussed so far as

$$\hat{\Sigma}_h = \mathbf{S} \hat{\Sigma}_{r,\text{ML}} \mathbf{S}^\top + \frac{1}{T} (\tilde{\mathbf{Y}} \mathbf{S}^\top - \hat{\mathbf{Y}})^\top (\tilde{\mathbf{Y}} \mathbf{S}^\top - \hat{\mathbf{Y}}), \quad (28)$$

where the first term is the residual sum of squares (of the reconciled forecast errors) and the second term is the squared difference between the base and reconciled forecast. Meanwhile, (27b) stresses separation of variation by a series of orthogonal projections of the base-forecast error similar to Cochran's theorem (see, e.g., Madsen and Thyregod, 2011).

The next theorem shows that as a consequence of the coherent projection, distance-reducing properties hold across any arbitrary splitting of the column space of the base-forecast errors.

Theorem 2.3 (Distance-reducing properties). A consequence of Lemma 2.1 is that for any column (and row) selection, $I \subseteq \{1, 2, \dots, n\}$, of $\hat{\mathbf{Y}}$ (and $\mathbf{S}$) then

$$\mathbf{S}_I (\mathbf{Y} - \tilde{\mathbf{Y}})^\top (\mathbf{Y} - \tilde{\mathbf{Y}}) \mathbf{S}_I^\top \leq (\mathbf{Y} \mathbf{S}_I^\top - \hat{\mathbf{Y}}_I)^\top (\mathbf{Y} \mathbf{S}_I^\top - \hat{\mathbf{Y}}_I), \quad (29)$$

where " $\leq$ " should be understood as " $rhs - lhs$ " being positive semi-definite. Furthermore, with $\mathbf{Y} \mathbf{S}_I^\top - \hat{\mathbf{Y}}_I \in \mathbb{R}^{T \times q_I}$, $\mathbf{y}_I = \text{vec}(\mathbf{S}_I \mathbf{Y}^\top)$, $\hat{\mathbf{y}}_I = \text{vec}(\mathbf{S}_I \hat{\mathbf{Y}}^\top)$, $\tilde{\mathbf{y}}_I = \text{vec}(\mathbf{S}_I \tilde{\mathbf{Y}}^\top)$, and $\mathbf{y}_I - \hat{\mathbf{y}}_I \in \mathcal{E}_I \subseteq$

$\mathbb{R}^{Tq_I}$, model (21) defines an orthogonal projection from $\mathcal{E}_I$ into a $q_I(n - m)$ -dimensional subspace where

$$\|\mathbf{y}_I - \hat{\mathbf{y}}_I\|^2 = \|\mathbf{y}_I - \tilde{\mathbf{y}}_I\|^2 + \|\tilde{\mathbf{y}}_I - \hat{\mathbf{y}}_I\|^2, \quad (30)$$

and, therefore,

$$\|\mathbf{y}_I - \tilde{\mathbf{y}}_I\|^2 \leq \|\mathbf{y}_I - \hat{\mathbf{y}}_I\|^2. \quad (31)$$

Proof. See Appendix G. ■

The result (31) is similar to the results presented in Panagiotelis et al. (2021, Theorems 3.1 and 3.2), but formulated for all observations in the training set rather than for an individual forecast. Notice that (29) generalises the result to variance–covariance matrices and shows that the distance reduction applies to any columns selection of $\mathbf{Y}\mathbf{S}^\top - \hat{\mathbf{Y}}$. Theorem 2.3 shows that the distance between the observations and the reconciled forecasts is less than the distance between the observations and the bottom-level base forecast, while Lemma 2.1 provides the separation of variation into orthogonal sums that is the basis for the central variance–covariance estimator and statistical tests.

2.5. Central covariance estimator

In the LM, series of orthogonal projections are used to construct hypothesis tests based on Cochran’s theorem, which leads to the F-test based on χ^2 -distributions and the central estimator for the variance. Since we are working with variance–covariance matrices, we need results related to the Whishart distribution (multivariate generalisation of the χ^2 -distribution).

Using the general definition of the Whishart distribution (see Lemma B.1, in Appendix B), we can derive the distribution of the sum of squares of each term in (27b), as given in the next corollary. In addition to defining the REML estimator as the central estimator for Σ_r , the corollary also points to a test for total homogeneity (see Section 4.1).

Corollary 2.2 (Central estimator for $\hat{\Sigma}_r$). Assuming model (21) and defining $\mathbf{H}$ as in Lemma 2.1, then

$$\left(\mathbf{Y} - \tilde{\mathbf{Y}}\right)^\top \left(\mathbf{Y} - \tilde{\mathbf{Y}}\right) \sim W_m(\Sigma_r, T - (n - m)), \quad (32)$$

implying that $\hat{\Sigma}_{r,\text{REML}}$ (23) is the central estimator for Σ_r . If, in addition $\beta_{\text{top}} = \mathbf{0}$, then

$$(\tilde{\mathbf{Y}} - \hat{\mathbf{Y}})^\top (\tilde{\mathbf{Y}} - \hat{\mathbf{Y}}) \sim W_m(\Sigma_r, n - m). \quad (33)$$

Proof. See Appendix H ■

3. Shrinkage and MAP estimation

In most realistic settings, there is a large number of parameters in the base-forecast variance–covariance matrix $\hat{\Sigma}_h$. The results so far use the estimated variance–covariance matrix directly. This is a disadvantage as it is well known that shrinkage is needed in order to stabilise the reconciliation weights (see, e.g., Nystrup et al., 2020). When applying shrinkage, the projections presented in Lemma 2.1 are no longer orthogonal. In this section, we show the equivalence between MAP estimation and forecast reconciliation when shrinkage is applied to the variance–covariance matrix Σ_h in (3).

In the following some result are formulated in terms of the marginal base-forecast variances divided into top and bottom levels, i.e.,

$$\hat{\Sigma}_h^{\text{d}} = \begin{bmatrix} \hat{\Sigma}_{h,\text{top}}^{\text{d}} & \mathbf{0} \\ \mathbf{0} & \hat{\Sigma}_{h,\text{bot}}^{\text{d}} \end{bmatrix}, \quad (34)$$

where both $\hat{\Sigma}_{h,\text{top}}^{\text{d}} \in \mathbb{R}^{(n-m) \times (n-m)}$ and $\hat{\Sigma}_{h,\text{bot}}^{\text{d}} \in \mathbb{R}^{m \times m}$ are diagonal matrices.

The usual shrinkage estimator for the variance–covariance matrix is

$$\hat{\Sigma}_s = (1 - \lambda)\hat{\Sigma}_h + \lambda\hat{\Sigma}_h^{\text{d}}, \quad (35)$$

with $\hat{\Sigma}_h = \frac{1}{T}(\mathbf{Y}\mathbf{S}^\top - \hat{\mathbf{Y}})^\top(\mathbf{Y}\mathbf{S}^\top - \hat{\mathbf{Y}})$ and $\hat{\Sigma}_h^{\text{d}} = \text{diag}(\hat{\Sigma}_h)$ (Wickramasuriya et al., 2019). The shrinkage parameter λ ensures that the matrix $\hat{\Sigma}_s$ is not ill-conditioned, but introduces bias in the estimation. Thus, (35) expresses a bias–variance tradeoff. The weight matrix calculated using $\hat{\Sigma}_s$ will be denoted

$$\mathbf{P}(\lambda) = (\mathbf{S}^\top \hat{\Sigma}_s^{-1} \mathbf{S})^{-1} \mathbf{S}^\top \hat{\Sigma}_s^{-1}. \quad (36)$$

In this section, we use MAP estimation. We assume that the conditional density for the observations is $f(\mathbf{y}|\boldsymbol{\theta})$ and that the parameters $\boldsymbol{\theta}$ follow some prior distribution $g(\boldsymbol{\theta})$. The MAP estimate is obtained as

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg \max_{\boldsymbol{\theta}} f(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta}). \quad (37)$$

The purpose is to show equivalence between the shrinkage estimator and an appropriate choice of prior distribution. The MAP formulation excludes estimation of parameters related to the prior distribution, which in our case would be λ . For fixed λ , the following theorem applies.

Theorem 3.1 (MAP and shrinkage). Assume the linear regression model

$$\mathbf{y}_t - \hat{\mathbf{y}}_{B,t} = [\mathbf{I}_m \otimes (\hat{\mathbf{y}}_{\text{top},t}^\top - \hat{\mathbf{y}}_{\text{bot},t}^\top \mathbf{S}_{\text{top}}^\top)] \boldsymbol{\beta}_{\text{top}} + \boldsymbol{\epsilon}_t; \quad \boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma}_r), \quad (38)$$

with prior distribution $\boldsymbol{\beta}_{\text{top}} \sim N(\boldsymbol{\beta}_{0,\text{top}}, \boldsymbol{\Sigma}_r \otimes \boldsymbol{\Sigma}_{\beta,0})$ and

$$\begin{aligned} \boldsymbol{\beta}_{0,\text{top}} &= \text{vec} \left[\left(\hat{\boldsymbol{\Sigma}}_{h,\text{top}}^{\text{d}} + \mathbf{S}_{\text{top}} \hat{\boldsymbol{\Sigma}}_{h,\text{bot}}^{\text{d}} \mathbf{S}_{\text{top}}^\top \right)^{-1} \mathbf{S}_{\text{top}} \hat{\boldsymbol{\Sigma}}_{h,\text{bot}}^{\text{d}} \right] \\ \boldsymbol{\Sigma}_{\beta,0} &= \frac{1-\lambda}{\lambda T} \left(\hat{\boldsymbol{\Sigma}}_{h,\text{top}}^{\text{d}} + \mathbf{S}_{\text{top}} \hat{\boldsymbol{\Sigma}}_{h,\text{bot}}^{\text{d}} \mathbf{S}_{\text{top}}^\top \right)^{-1}. \end{aligned} \quad (39)$$

With the choice of design matrix given in (38), the MAP estimate of $\boldsymbol{\beta}$ is independent of $\boldsymbol{\Sigma}_r$ and can be written as (with $\hat{\boldsymbol{\beta}}_{\text{top}} = \text{vec}(\hat{\boldsymbol{\beta}}_{\text{top}}^m)$)

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\text{top}}^m &= \left((1-\lambda) \left(\hat{\mathbf{Y}}_{\text{top}} - \hat{\mathbf{Y}}_{\text{bot}} \mathbf{S}_{\text{top}}^\top \right)^\top \left(\hat{\mathbf{Y}}_{\text{top}} - \hat{\mathbf{Y}}_{\text{bot}} \mathbf{S}_{\text{top}}^\top \right) + \lambda T \left(\hat{\boldsymbol{\Sigma}}_{h,\text{top}}^{\text{d}} + \mathbf{S}_{\text{top}} \hat{\boldsymbol{\Sigma}}_{h,\text{bot}}^{\text{d}} \mathbf{S}_{\text{top}}^\top \right) \right)^{-1} \times \\ &\quad \left((1-\lambda) \left(\hat{\mathbf{Y}}_{\text{top}} - \hat{\mathbf{Y}}_{\text{bot}} \mathbf{S}_{\text{top}}^\top \right)^\top \left(\mathbf{Y} - \hat{\mathbf{Y}}_{\text{bot}} \right) + \lambda T \mathbf{S}_{\text{top}} \hat{\boldsymbol{\Sigma}}_{h,\text{bot}}^{\text{d}} \right). \end{aligned} \quad (40)$$

where $\hat{\mathbf{Y}}_{\text{top}} = \text{vec}^{-1}(\hat{\mathbf{y}}_{\text{top}}) \in \mathbb{R}^{T \times (n-m)}$ and $\hat{\mathbf{Y}}_{\text{bot}} = \text{vec}^{-1}(\hat{\mathbf{y}}_{\text{bot}}) \in \mathbb{R}^{T \times m}$ are the collection of top- and bottom-level forecasts, respectively. The solution is equivalent to $\mathbf{P}(\lambda)$ (see (36)) in the sense that

$$\mathbf{P}(\lambda)^\top = \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} + \begin{bmatrix} \mathbf{I} \\ -\mathbf{S}_{\text{top}}^\top \end{bmatrix} \hat{\boldsymbol{\beta}}_{\text{top}}^m. \quad (41)$$

Proof. Relies on results presented in Wickramasuriya et al. (2019, Theorem 1). See Appendix I for details. ■

The problem defined by (38)–(39) is equivalent to the generalized Tikhonov regularisation (see, e.g., Kalivas, 2012); that is, the solution to the minimisation problem

$$\arg \min_{\beta_{\text{top}}} \left(\|\mathbf{X}\beta_{\text{top}} - (\mathbf{y} - \hat{\mathbf{y}}_{\text{bot}})\|_{\Sigma^{-1}}^2 + \|\beta_{\text{top}} - \beta_{0,\text{top}}\|_{\Sigma_{\beta}^{-1}}^2 \right) \quad (42)$$

with $\mathbf{X}$ defined by (38), $\Sigma = \mathbf{I}_T \otimes \Sigma_r$ and the shrinkage target defined by (39). Other targets could be chosen. For example, the usual Ridge regression ($\beta_{0,\text{top}} = \mathbf{0}$ and $\Sigma_{\beta} = \frac{1}{\lambda} \mathbf{I}$, ($\lambda > 0$)) would define the bottom-level base-forecast error variance–covariance as the shrinkage target.

The next corollary and theorem state the MAP estimate of the variance for a given λ and a prior for Σ_r such that it is equivalent to the result given in Wickramasuriya et al. (2019, Lemma 1).

Corollary 3.1 (MAP estimate of Σ_r). Using the model in Theorem 3.1, the MAP estimate of Σ_r is

$$\hat{\Sigma}_{r,\text{MAP}} = \frac{T/(1-\lambda)}{T+n-m} \left(\mathbf{P}(\lambda) \hat{\Sigma}_s \mathbf{P}^{\top}(\lambda) - \lambda \left(\left(\hat{\Sigma}_{h,\text{bot}}^{\text{d}} \right)^{-1} + \mathbf{S}_{\text{top}}^{\top} \left(\hat{\Sigma}_{h,\text{top}}^{\text{d}} \right)^{-1} \mathbf{S}_{\text{top}} \right)^{-1} \right). \quad (43)$$

Proof. See Appendix J. ■

The result in Corollary 3.1 shows that the difference in variance depends on λ through a multiplicative factor and on the shrinkage target (prior for the weight matrix). For illustrative purposes consider $\lambda = \frac{n-m}{T+n-m}$, then the difference is

$$\mathbf{P}(\lambda) \hat{\Sigma}_s \mathbf{P}^{\top}(\lambda) - \hat{\Sigma}_{r,\text{MAP}} = \lambda \left(\left(\hat{\Sigma}_{h,\text{bot}}^{\text{d}} \right)^{-1} + \mathbf{S}_{\text{top}}^{\top} \left(\hat{\Sigma}_{h,\text{top}}^{\text{d}} \right)^{-1} \mathbf{S}_{\text{top}} \right)^{-1}. \quad (44)$$

Now consider the limit $\lambda \rightarrow 1$. Using (24) we get

$$\hat{\mathbf{V}}_1 = \lim_{\lambda \rightarrow 1} \hat{\mathbf{V}}[\mathbf{y}_t - \tilde{\mathbf{y}}_t] = \mathbf{P}(1) \hat{\Sigma}_h^{\text{d}} \mathbf{P}(1)^{\top} = \left(\left(\hat{\Sigma}_{h,\text{bot}}^{\text{d}} \right)^{-1} + \mathbf{S}_{\text{top}}^{\top} \left(\hat{\Sigma}_{h,\text{top}}^{\text{d}} \right)^{-1} \mathbf{S}_{\text{top}} \right)^{-1}, \quad (45)$$

while (43) becomes

$$\begin{aligned} \lim_{\lambda \rightarrow 1} \hat{\Sigma}_{r,\text{MAP}} &= \frac{T}{T+n-m} \mathbf{P}(1) \hat{\Sigma}_h \mathbf{P}(1)^{\top} \\ &= \frac{T}{T+n-m} \mathbf{V}_1 \mathbf{S}_{\text{top}}^{\top} \left(\hat{\Sigma}_h^{\text{d}} \right)^{-1/2} \mathbf{R}_h \left(\hat{\Sigma}_h^{\text{d}} \right)^{-1/2} \mathbf{S}_{\text{top}} \mathbf{V}_1, \end{aligned} \quad (46)$$

where $\mathbf{R}_h$ is the correlation matrix corresponding to $\hat{\Sigma}_h$ (see K.1 for details on showing (45) and (46)). Hence, the estimator $\hat{\Sigma}_{r,\text{MAP}}$ includes the correlation of the base-forecast errors, even when it is excluded from the estimation of mean-value parameters. Finally, (43) can be written as

$$\hat{\Sigma}_{r,\text{MAP}} = \frac{T}{T+n-m} \left(\mathbf{P}(\lambda) \hat{\Sigma}_h \mathbf{P}^{\top}(\lambda) + \frac{\lambda}{1-\lambda} \left(\mathbf{P}(\lambda) \hat{\Sigma}_h^{\text{d}} \mathbf{P}^{\top}(\lambda) - \mathbf{P}(1) \hat{\Sigma}_h^{\text{d}} \mathbf{P}^{\top}(1) \right) \right). \quad (47)$$

The following theorem introduces a prior for the variance that will give a variance estimate equivalent to the usual forecast reconciliation with shrinkage.

Theorem 3.2 (Prior for Σ_r). If, in addition to the prior given in Theorem 3.1, the prior for the variance–covariance matrix Σ_r is chosen as the inverse Wishart $\Sigma_r \sim \mathcal{W}^{-1}(\Psi, v)$ ($v > m - 1$) with

$$\Psi = \frac{\lambda T}{1 - \lambda} \left(\left(\hat{\Sigma}_{h,\text{bot}}^{\text{d}} \right)^{-1} + \mathbf{S}_{\text{top}}^{\top} \left(\hat{\Sigma}_{h,\text{top}}^{\text{d}} \right)^{-1} \mathbf{S}_{\text{top}} \right)^{-1}, \quad (48)$$

and $v = \frac{\lambda T}{1 - \lambda} - (n + 1)$ (implying that $\lambda > \frac{m+n}{T+m+n}$), then the MAP estimate of $\hat{\Sigma}_r$ is

$$\hat{\Sigma}_{r,\text{shrink}} = \mathbf{P}(\lambda) \hat{\Sigma}_h \mathbf{P}^{\top}(\lambda). \quad (49)$$

Proof. See Appendix K. ■

Note that if $\lambda \leq \frac{m+n}{T+m+n}$ ($v \leq m - 1$), then the probability density function of the inverse Wishart is not defined and the MAP interpretation of the above is not valid. However, for any fixed $\lambda > 0$, implying that $v > -(n + 1)$, (K.1) in Appendix K defines the solution (49) in a regularisation setting.

In Theorem 3.2, the prior for the forecast variance is chosen independently of the prior for the mean. It is possible to choose other priors for Σ_r depending on the specific application, e.g., independence between observations, AR(1) correlation structure, or not to have any prior for the residual variance. The choice above simply shows the equivalence with the formulation of the forecast variance given by Wickramasuriya et al. (2019). In the formulation of Theorems 3.1 and 3.2, λ is the only hyperparameter.

In the previous derivation we have included REML estimates for each of the models. A similar result for the model presented in Theorems 3.1 and 3.2 is given in the corollary below.

Corollary 3.2 (REML and MAP). Using the model in Theorems 3.1–3.2, the REML correction term (defined as the log-determinant of the Hessian of the log-likelihood wrt. β) is

$$\log |\mathbf{X}^{\top} \Sigma^{-1} \mathbf{X} + \Sigma_{\beta}^{-1}|. \quad (50)$$

If $\mathbf{x}_{i,t} = \mathbf{x}_{j,t} \in \mathbb{R}^{n-m}$, then the MAP–REML estimate of Σ_r is

$$\hat{\Sigma}_{r,\text{sREML}} = \frac{T}{T - (n - m)(1 - \lambda)} \mathbf{P}(\lambda) \hat{\Sigma}_s \mathbf{P}^{\top}(\lambda). \quad (51)$$

Proof. See Appendix L. ■

In the general case where $\mathbf{x}_{i,t} \neq \mathbf{x}_{j,t}$, the MAP-REML estimate of Σ_r is more complicated since the priors should be recalculated. The obvious application would be that some columns of $\mathbf{X}$ are removed due to non-significance, in which case the prior needs to be recalculated and the derivative of the determinant wrt. Σ_r should be recalculated in line with the result in (17). We will not explore this further here.

The methods used in this section imply that parameter and forecast variance can be written explicitly on the form that we discuss in the next section.

4. Parameter and forecast variance

Using the general theory for LMs (univariate and iid. residuals), the variance of a forecast is given by

$$\hat{V}[y_{t+h}] = \mathbf{x}_{t+h}^\top \hat{V}[\hat{\beta}] \mathbf{x}_{t+h} + \hat{\sigma}^2. \quad (52)$$

In the cases covered in this article, (52) can be formulated as

$$\hat{V}[\mathbf{y}_{t+h}] = \hat{V}[\mathbf{X}_{\cdot, t+h} \hat{\beta} + \boldsymbol{\epsilon}_{t+h}] = \mathbf{X}_{\cdot, t+h} \hat{V}[\hat{\beta}] \mathbf{X}_{\cdot, t+h}^\top + \hat{\Sigma}_r. \quad (53)$$

The first term relates to the uncertainty of the estimated parameters (referred to as reconciliation-matrix estimation error by Pritularga et al. (2021)). The second term relates to the usual stochastic uncertainty (referred to as the irreducible error in Panagiotelis et al. (2021)).

The exact form of $\hat{V}[\hat{\beta}]$ depends on the model formulation, but in the general case it is

$$\hat{V}[\hat{\beta}] = \left(\mathbf{X}^\top \left(\mathbf{I}_T \otimes \hat{\Sigma}_r^{-1} \right) \mathbf{X} \right)^{-1}. \quad (54)$$

In case $\mathbf{X}_{\cdot, t} = \mathbf{I}_m \otimes \mathbf{x}_t^\top$, this simplifies to

$$\hat{V}[\hat{\beta}] = \hat{\Sigma}_r \otimes \left(\mathbf{X}_{1,\cdot}^\top \mathbf{X}_{1,\cdot} \right)^{-1}. \quad (55)$$

Any of the estimates of $\hat{\Sigma}_r$ can be applied (usually the REML estimate would be preferred). When (53) is used with $\hat{\Sigma}_{r, \text{sREML}}$, it will be referred to as

$$\hat{\Sigma}_{r, \text{par}}(\mathbf{X}_{\cdot, t+h}) = \mathbf{X}_{\cdot, t+h} \hat{V}[\hat{\beta}] \mathbf{X}_{\cdot, t+h}^\top + \hat{\Sigma}_{r, \text{sREML}}, \quad (56)$$

and we will use the short-hand notation $\hat{\Sigma}_{r,\text{par}} := \hat{\Sigma}_{r,\text{par}}(\mathbf{X}_{\cdot,t+h})$.

In the presence of prior distributions, we can write the MAP estimator (40) as an affine transformation of the ML estimate

$$\hat{\beta}_{\text{top}} = \mathbf{W}_{1,\lambda} \mathbf{W}_2 \hat{\beta}_{\text{top,ML}} + \mathbf{W}_{1,\lambda} \Sigma_{\beta}^{-1} \beta_{0,\text{top}}, \quad (57)$$

where $\mathbf{W}_{1,\lambda} = (\mathbf{X}^{\top}(\mathbf{I}_T \otimes \Sigma_r)^{-1} \mathbf{X} + \Sigma_{\beta}^{-1})^{-1}$, $\mathbf{W}_2 = \mathbf{X}^{\top}(\mathbf{I}_T \otimes \Sigma_r)^{-1} \mathbf{X}$, and $\hat{\beta}_{\text{top,ML}} = (\mathbf{X}^{\top}(\mathbf{I}_T \otimes \Sigma_r)^{-1} \mathbf{X})^{-1} \mathbf{X}^{\top}(\mathbf{I}_T \otimes \Sigma_r)^{-1}(\mathbf{y} - \hat{\mathbf{y}}_{\text{bot}})$ is the ML estimator of $\hat{\beta}_{\text{top}}$ (van Wieringen, 2023). Ignoring the variance of the priors (i.e., $\hat{\Sigma}_h^{\text{d}}$), the variance of $\hat{\beta}_{\text{top}}$ is

$$\begin{aligned} \hat{V}[\hat{\beta}_{\text{top}}] &= \mathbf{W}_{1,\lambda} \mathbf{W}_2 V[\hat{\beta}_{\text{top,ML}}] \mathbf{W}_2 \mathbf{W}_{1,\lambda} \\ &= \hat{\Sigma}_r \otimes \left((\mathbf{X}_{1,\cdot}^{\top} \mathbf{X}_{1,\cdot} + \Sigma_{\beta,0})^{-1} \mathbf{X}_{1,\cdot}^{\top} \mathbf{X}_{1,\cdot} (\mathbf{X}_{1,\cdot}^{\top} \mathbf{X}_{1,\cdot} + \Sigma_{\beta,0})^{-1} \right). \end{aligned} \quad (58)$$

Using (41) and (B.6) we can find the variance–covariance of the weight matrix as

$$\begin{aligned} \hat{V} \left[\text{vec} \left(\mathbf{P}^{\top} \right) \right] &= \hat{V} \left[\text{vec} \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} - \begin{bmatrix} \mathbf{I} \\ \mathbf{S}_{\text{top}}^{\top} \end{bmatrix} \hat{\beta}_{\text{top}}^m \right) \right] \\ &= \left(\mathbf{I}_m \otimes \begin{bmatrix} \mathbf{I} \\ \mathbf{S}_{\text{top}}^{\top} \end{bmatrix} \right) \hat{V} [\hat{\beta}_{\text{top}}] \left(\mathbf{I}_m \otimes \begin{bmatrix} \mathbf{I} \\ \mathbf{S}_{\text{top}}^{\top} \end{bmatrix} \right)^{\top}. \end{aligned} \quad (59)$$

This can be used for calculating standard errors of the weight matrix and pairwise correlations. The standard errors are the basis for the usual Wald test.

4.1. Model reduction

The theory introduced in the previous sections supports general test strategies, such as ANOVA-type (likelihood-ratio) tests for specific hypothesis and partial (Wald) tests for specific parameters. The latter supports an exploratory approach where the least significant parameters are removed one by one.

The obvious null hypothesis is that some parameters could be equal to zero. Given the parameter variance, it is straight forward to construct the Wald test–statistics as

$$z_{\text{obs},i} = \frac{\hat{\beta}_{\text{top},i}}{\sqrt{\hat{V}[\hat{\beta}_{\text{top}}]_{ii}}}. \quad (60)$$

This can be compared to a Student t -distribution with appropriate number of degrees of freedom, which can be approximated by a standard normal in most realistic examples.

In light of Lemma 2.1, Lemma B.1, and Corollary 2.2, a reasonable test statistic for the null hypothesis $\beta = \mathbf{0}$ against the alternative would be

$$F_I = \frac{\mathbf{1}^\top (\tilde{\mathbf{Y}} \mathbf{S}_I^\top - \hat{\mathbf{Y}}_I)^\top (\tilde{\mathbf{Y}} \mathbf{S}_I^\top - \hat{\mathbf{Y}}_I) \mathbf{1} / (n - m)}{\mathbf{1}^\top \mathbf{S}_I (\mathbf{Y} - \tilde{\mathbf{Y}})^\top (\mathbf{Y} - \tilde{\mathbf{Y}}) \mathbf{S}_I^\top \mathbf{1} / (T - n + m)} = \frac{\sum_{i,j \in I} \mathbf{S} \mathbf{S}_{\text{mod},ij} / (n - m)}{\sum_{i,j \in I} \mathbf{S} \mathbf{S} \mathbf{E}_{ij} / (T - n + m)}, \quad (61)$$

where the second equality is notation and F_I should be compared to an F -distribution with $n - m$ and $T - n + m$ degrees of freedom. When shrinkage is applied, the parameter variance will depend on the shrinkage parameter. We will not explore this point further.

5. Simulation study

The purpose of this simulation study is to compare different estimators for the forecast-error variance–covariance matrix and discuss the impact of parameter uncertainty. We simulate the data-generating process at the bottom level ($\mathbf{y}_t \in \mathbb{R}^4$) using a multivariate AR(1) process

$$\mathbf{y}_t = \mathbf{A} \mathbf{y}_{t-1} + \boldsymbol{\epsilon}_t; \quad \boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon) \quad \text{and iid.}, \quad (62)$$

with $\mathbf{A}_{ii} = 0.6$, $\mathbf{A}_{ij} = 0.1$ ($i \neq j$), and $\boldsymbol{\Sigma}_\epsilon = \mathbf{I}$.

The simulated bottom level is aggregated using the summation matrix (2). Independent, univariate AR(1) models are estimated on each level and the forecast errors are calculated based on the estimated models.

As an example, assume that the diagonal elements of the error variance–covariance matrix are estimated as $\hat{\boldsymbol{\Sigma}}_h^d = [4, 2, 2, 1, 1, 1, 1]$, which is not far from the observed values. In that case the prior weight matrix (calculated from $\beta_{0,\text{top}}$ (39) and the linear constraints (4)) is

$$\mathbf{P}(1) = \begin{bmatrix} 0.09 & 0.20 & -0.04 & 0.72 & -0.28 & -0.05 & -0.05 \\ 0.09 & 0.20 & -0.04 & -0.28 & 0.72 & -0.05 & -0.05 \\ 0.09 & -0.04 & 0.20 & -0.05 & -0.05 & 0.72 & -0.28 \\ 0.09 & -0.04 & 0.20 & -0.05 & -0.05 & -0.28 & 0.72 \end{bmatrix}, \quad (63)$$

where the first three columns of $P(1)$ correspond to $\beta_{0,\text{top}}$. The prior variance of β_{top} (see (39)) is

$$\Sigma_\beta = \Sigma_r \otimes \begin{bmatrix} 0.16 & -0.08 & -0.08 \\ -0.08 & 0.28 & 0.04 \\ -0.08 & 0.04 & 0.28 \end{bmatrix}, \quad (64)$$

and the prior for the variance Σ_r (see (48)) is

$$\Psi = \begin{bmatrix} 0.72 & -0.28 & -0.05 & -0.05 \\ -0.28 & 0.72 & -0.05 & -0.05 \\ -0.05 & -0.05 & 0.72 & -0.28 \\ -0.05 & -0.05 & -0.28 & 0.72 \end{bmatrix}. \quad (65)$$

Considering the priors and the prior variance, it is reasonable to assume that statistical tests would lead to $P_{1,3}$, $P_{2,3}$, $P_{3,2}$, $P_{4,2}$, and possibly parameters from the first column being removed before other parameters. The prior, Ψ , for Σ_r introduces a covariance structure that dominates in small-sample cases (λ large) through the (arbitrary) structure of S and the diagonal elements of $\hat{\Sigma}_h$, the structure of S is arbitrary in the sense that we could have chosen other aggregation levels in the modelling. In our simulation example, a better prior would be $\Psi = \sigma_0^2 I$.

We test three different estimates of the forecast variance:

1. The usual MAP shrinkage estimate $\hat{V}[\mathbf{y}_{t+1}] = \hat{\Sigma}_{r,\text{shrink}}$ (see Theorem 3.2).
2. The MAP-REML shrinkage estimate $\hat{V}[\mathbf{y}_{t+1}] = \hat{\Sigma}_{r,\text{sREML}}$ (see Corollary 3.2).
3. The forecast variance $\hat{\Sigma}_{r,\text{par}}$ defined by (56) using the MAP-REML shrinkage estimate of Σ_r .

In all three cases we use the optimal value for λ_{opt} given by Schäfer and Strimmer (2005) and implemented in the R-package `corpcor` (Schafer et al., 2021).

The quality of the forecasts is evaluated out of sample using the log-score (Gneiting and Raftery, 2007; Bjerregård et al., 2021) for the bottom level

$$\text{LogS}(\mathbf{y}, \tilde{\mathbf{y}}, \hat{\Sigma}_r) = - \sum_{t=1}^N \log(\phi(\mathbf{y}_{t+1}; \tilde{\mathbf{y}}_{t+1}, \hat{V}[\mathbf{y}_{t+1}])), \quad (66)$$

where $\phi(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the density of the multivariate normal with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$. The relative log-score is the difference between the log-score of two different models.

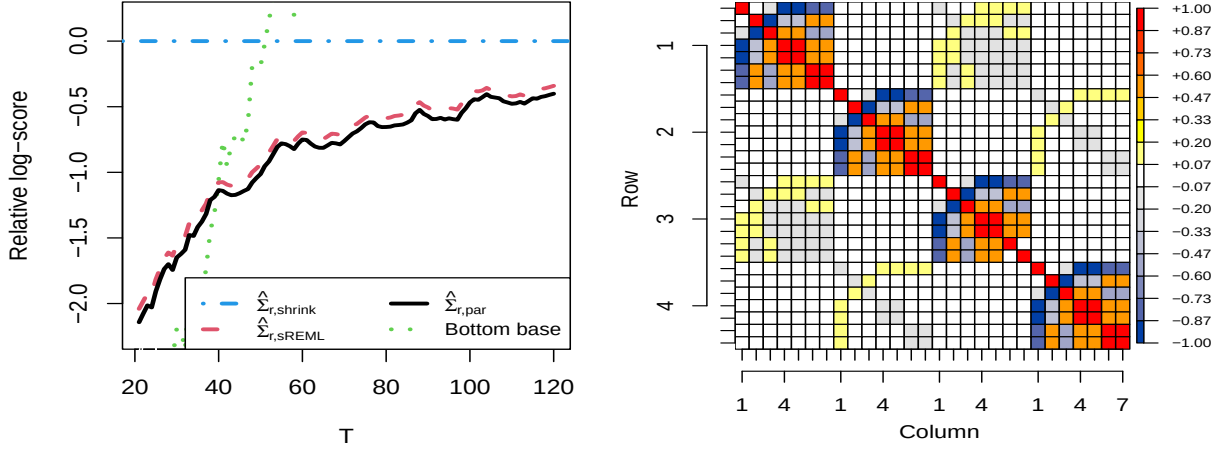


Figure 1: Left: Relative (to $\hat{\Sigma}_{r,shrink}$) log-score for different estimates of the forecast variance. Lines indicate average relative log-score (over $\{T-4, \dots, T+4\}$). Right: Example of the parameter correlation matrix ($T = 120$). Rows and columns refer to the weight matrix (63).

We simulate 50 times for each length of training and test period. The result is shown in Figure 1. We see that the base forecasts perform better than all three methods for small samples (up to about 40 observations). There is a large and consistent gain in using the REML estimator for Σ_r . The forecast variance including the parameter uncertainty consistently produces better results than the other methods, although the difference is small. The effect of the choice of variance model becomes smaller as the number of observations in the training set increases. This is not surprising as the REML correction gets smaller with increasing sample size, as does the correction for parameter uncertainty.

For illustration purposes, we present the result for one of the realisations behind the results in Figure 1. The result is for $T = 120$ using the optimal value for λ , which in this case was $\lambda_{opt} = 0.056$. The estimated weight matrix is

$$\mathbf{P}(\lambda_{opt}) = \begin{bmatrix} 0.14 & 0.13 & 0.03 & 0.72 & -0.28 & -0.18 & -0.18 \\ 0.08 & 0.16 & -0.01 & -0.24 & 0.76 & -0.07 & -0.07 \\ 0.19 & -0.01 & 0.14 & -0.18 & -0.18 & 0.66 & -0.34 \\ 0.22 & 0.02 & 0.01 & -0.24 & -0.24 & -0.23 & 0.77 \end{bmatrix}. \quad (67)$$

The REML estimate of Σ_r is

$$\hat{\Sigma}_{r,s\text{REML}} = \begin{bmatrix} 1.23 & 0.14 & -0.15 & -0.07 \\ 0.14 & 1.11 & -0.05 & -0.02 \\ -0.15 & -0.05 & 1.03 & -0.14 \\ -0.07 & -0.02 & -0.14 & 0.99 \end{bmatrix} \quad (68)$$

and the standard errors related to $P(\lambda)$ are

$$se(P(\lambda_{\text{opt}})) = \begin{bmatrix} 0.075 & 0.078 & 0.025 & 0.088 & 0.064 & 0.085 & 0.070 \\ 0.026 & 0.093 & 0.066 & 0.088 & 0.071 & 0.067 & 0.070 \\ 0.070 & 0.093 & 0.074 & 0.069 & 0.071 & 0.023 & 0.084 \\ 0.078 & 0.071 & 0.074 & 0.024 & 0.085 & 0.063 & 0.084 \end{bmatrix}. \quad (69)$$

Only the first three columns of the weight matrix are actual parameters. The rest are determined by linear constraints. We see that only three parameters (from the first column) are more than two standard errors away from zero. Hence, from a partial test perspective most parameters could be set to zero. As the linear constraints introduce correlation, the testing should be done stepwise by recalculating everything or consider the correlation when testing multiple parameters.

The correlation matrix corresponding to the parameters is presented in Figure 1. There is a high (negative) correlation between $P_{i,2}$ and $P_{i,3}$ (e.g., between $P_{1,2}$ and $P_{1,3}$) and there are high correlations between the top- and bottom-level weights (introduced by the linear constraints). In this example, there is a weak correlation between different rows in the weight matrix. This is due to the weak correlation in the estimated variance–covariance matrix (68).

6. Forecasting electricity load

As a case study, we consider forecasts of electricity load in Sweden. The load is divided into four areas plus the total (sum of the four areas). This gives a total of five series. The considered data spans a period of five years from 2016 to 2020. The years 2016–2019 are used for estimating a mean-value model including seasonal and diurnal variation. The 2019 residuals from that model are used to estimate double-seasonal AR-models with weekly and daily variation. Further details on the model and data can be found in Møller et al. (2024). We will only briefly outline the parts that are central for illustrating the methods presented in this article.

	$\frac{\ \mathbf{y}_I - \hat{\mathbf{y}}_I\ ^2}{T}$	$\frac{\ \mathbf{y}_I - \tilde{\mathbf{y}}_I\ ^2}{T}$	$\frac{\ \tilde{\mathbf{y}}_I - \hat{\mathbf{y}}_I\ ^2}{T}$	$\frac{\frac{\sum SS_{mod}}{n-m}}{\frac{\sum SSE}{T-n+m}}$	$\frac{\ \mathbf{y}_I - \tilde{\mathbf{y}}_I^\lambda\ ^2}{T}$	$\frac{\ \tilde{\mathbf{y}}_I^\lambda - \hat{\mathbf{y}}_I\ ^2}{T}$	$\frac{2(\tilde{\mathbf{e}}_I^\lambda)^T \tilde{\mathbf{e}}_I^\lambda}{T}$
24h	220.2	113.25	106.91	8.63	131.03	81.58	7.55
12h	101.3	67.31	33.95	4.85	78.09	20.71	2.46
8h	66.7	49.95	16.75	3.13	57.90	6.91	1.88
6h	52.9	39.88	13.00	2.96	46.23	5.07	1.58
4h	38.2	27.88	10.29	3.63	32.33	4.64	1.21
3h	29.8	21.52	8.27	3.85	24.91	3.89	0.99
2h	19.6	14.72	4.87	3.10	17.05	1.89	0.65
1h	10.6	7.63	2.97	3.64	8.83	1.42	0.34
Total	539.1	342.14	197.00	3.27	396.36	126.13	16.66

Table 1: Variance separation (30) on the training set for area SE. Column one is the sum of squared base-forecast errors on the training set. Columns two and three illustrate the orthogonal projection by adding up to column one. Column four is the test statistic (61). Columns five to seven show the results using shrinkage ($\tilde{\mathbf{e}}_I^\lambda = \mathbf{y}_I - \tilde{\mathbf{y}}_I^\lambda$ and $\tilde{\mathbf{e}}_I^\lambda = \tilde{\mathbf{y}}_I^\lambda - \hat{\mathbf{y}}_I$).

Base forecasts are made using AR models, as described above, once every 24 hours for the next 24 hours. The levels of the models are 1, 2, 3, 4, 6, 8, 12, and 24 hours. This implies that $n = 60$ and $m = 24$. The purpose is to illustrate the estimation methods for the base-forecast variance-covariance matrix, the parameter variance-covariance, and the orthogonal projections. We focus on the bottom level (hourly forecast 1–24 hours ahead) in the accuracy evaluation.

Table 1 illustrates variance separation (30) on the training set for area SE. Column one is the sum of squared base-forecast errors on the training set. Columns two and three illustrate the orthogonal projection by adding up to column one. Comparing column four of Table 1 to an F-distribution with $n - m = 36$ and $T - n + m = 365 - 36 = 329$ degrees of freedom (95%-quantile is 1.45) shows that the improvement in sum of squared errors is significant on all levels of the hierarchy, with the largest improvement occurring at the top level. When shrinkage is applied, the sums of squared errors increase and the projection is no longer orthogonal (last column is not equal to zero). Similar tables for the individual areas are given in N.

If the assumptions of the model were true, then the obvious metric for evaluation would be the log-score. As the log-score is very sensitive to deviations in the distribution assumption (Bjerregård

	SE	SE1	SE2	SE3	SE4
RMSE(base)	0.60	0.11	0.19	0.39	0.15
RRMSE(reconciled)	-5.78	-2.74	-6.31	-5.69	-8.93
LogS(base)	-615.11	-8361.17	-6761.02	-6297.84	-13808.00
relLogS($\hat{\Sigma}_{r, \text{shrink}}$)	-116.85	128.23	-322.59	197.63	251.12
relLogS($\hat{\Sigma}_{r, \text{sREML}}$)	-104.46	-246.21	-454.52	308.47	382.18
relLogS($\hat{\Sigma}_{r, \text{par}}$)	-102.82	-333.07	-470.30	321.75	393.28
Vs(base)	36779.81	118.98	361.77	6616.91	241.28
relVs($\hat{\Sigma}_{r, \text{shrink}}$)	-17.40	0.24	-1.52	-21.88	-17.56
relVs($\hat{\Sigma}_{r, \text{sREML}}$)	-16.56	-0.92	0.98	-20.22	-16.68
relVs($\hat{\Sigma}_{r, \text{par}}$)	-16.54	-1.23	1.36	-20.08	-16.69

Table 2: Summary statistics for out-of-sample load forecasts at the bottom level. The first two rows are the accuracy of the base and reconciled forecasts. The log-score evaluates the full distribution and the last rows evaluate the covariance estimates using the variogram score.

et al., 2021), we will also consider the variogram score (see, e.g., Scheuerer and Hamill, 2015)

$$Vs_p(F, \mathbf{y}) = \sum_{i,j}^m w_{ij} (|y_i - y_j|^p - E_F[|Y_i - Y_j|^p])^2, \quad (70)$$

where F is the forecast distribution, $\mathbf{y}$ is a set of observations, w_{ij} is a weight function, and p is the order of the variogram score. We choose $w_{ij} = 1$ and $p = 2$, implying that there is a closed-form solution for $E[|Y_i - Y_j|^p]$ under the Gaussian assumption (see Scheuerer and Hamill, 2015).

In addition to the variogram and log-score, we consider the relative root mean square error defined by

$$RRMSE = \frac{RMSE_{\text{Reconciled}} - RMSE_{\text{Base}}}{RMSE_{\text{Base}}} \cdot 100\%. \quad (71)$$

In a similar way, we calculate a relative variogram score as

$$relVs = \frac{Vs_2(F, \mathbf{y}) - Vs_2(F_{\text{Base}}, \mathbf{y})}{Vs_2(F_{\text{Base}}, \mathbf{y})} \cdot 100\%. \quad (72)$$

All evaluated models have the same mean-value prediction and, hence, also the same RRMSE. Using the observed bottom-level base-forecast error variance–covariance matrix as reference, we evaluate different estimators for the forecast variance–covariance of the reconciled forecast error. A summary of the results is shown in Table 2. RRMSE is improved in all areas. For most models

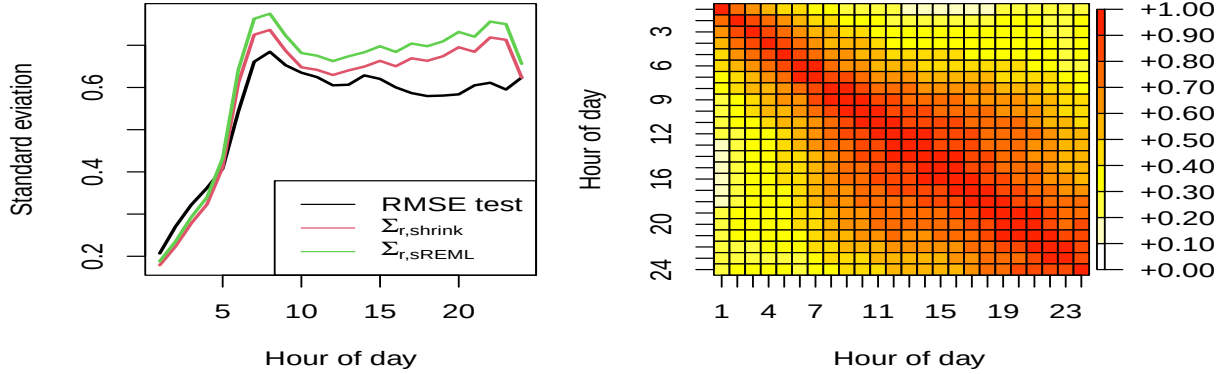


Figure 2: Left: Out-of-sample RMSE and standard deviation for different estimates of the variance–covariance matrix. Right: Correlation matrix corresponding to $\hat{\Sigma}_{r,sREML}$.

the log-score is very good in some areas and very bad in other areas, which shows the sensitivity of the log-score. The relative variogram score is more consistent across the models.

The estimated correlation matrix and standard deviation for the entire Sweden (SE) is shown in Figure 2. The autocorrelation and increasing variance in the residuals is very clear. In this particular case the observed variance on the test set is a bit smaller than the estimate. From similar plots for the other areas (Figure M.7) it is clear that it is not a general picture that the observed variance on the test set is smaller. In SE2 and SE4, the estimated variance on horizon 22–23 is very high, which is an indication of outliers.

The weights, their standard errors, and the ratio between the absolute values of the weights and their standard errors (absolute value of the Wald test statistic for the hypothesis that the weight is zero) are shown in Figure 3 for area SE (see Appendix M, Figure M.6 for other areas). It is difficult to distinguish clear patterns in the weights and standard errors; however, it is clear that the Wald test statistics are high for two-hour forecasts as explanatory variables for the two corresponding one-hour forecasts. There are many weights that from a testing perspective are candidates for being removed.

The correlation matrix for the first two rows and the first 36 columns in the weight matrix (first 72 of 864 elements of $\hat{\beta}_{top}$) is shown in Figure 4. It is clear that there is correlation between some weights from the same row and that the correlation is decreasing with the distance in time. The correlation between forecasts on different levels (e.g., weights related to two- and three-hour

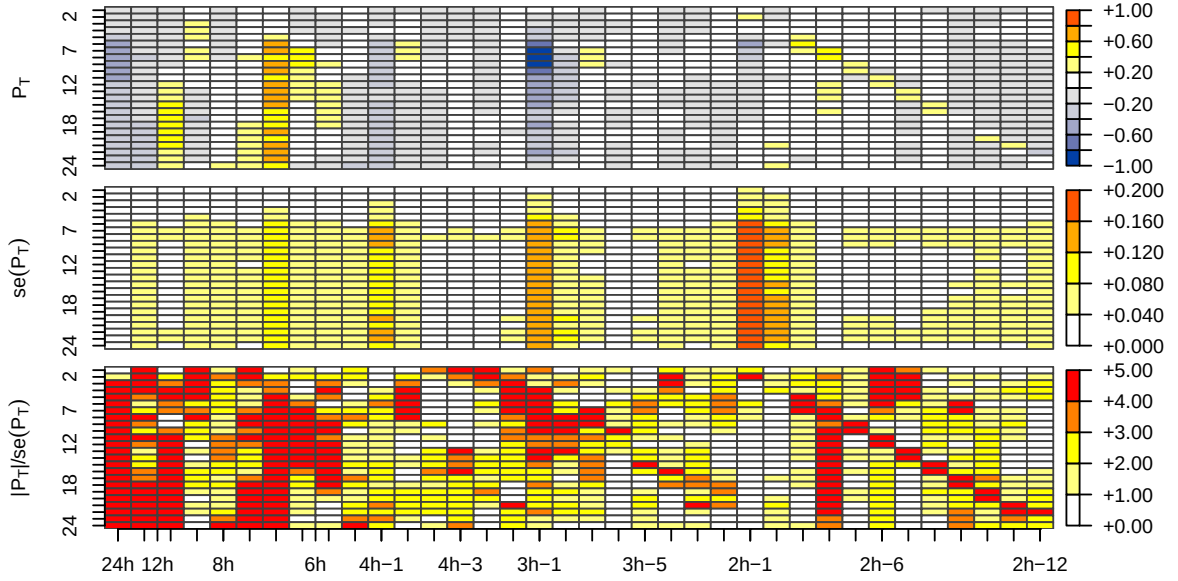


Figure 3: Top: Volume-weighted weight matrix. Middle: Corresponding standard errors. Bottom: Absolute value of the weights divided by standard errors.

forecasts) are generally low both within and between rows. Further, the correlation between row one and two is strong for the same predictor but decreases for rows further away (see Figure M.5 in M for the full correlation matrix). This decreasing correlation is a consequence of the decreasing correlation in the residual correlation matrix shown in Figure 2.

7. Conclusion

We have formulated the forecast reconciliation problem as a LM to enable uncertainty quantification and shown the equivalence between usual forecast reconciliation and the MAP estimates for specific choices of priors. This formulation allowed us to build on well-known results from regression analysis and formulate the REML estimate of the forecast-error variance–covariance matrix. The LM formulation implied that forecast reconciliation can be viewed as orthogonal projections for all levels of the hierarchy, which we used to prove in-sample distance-reducing properties.

A further benefit of the proposed LM formulation was the possibility of estimating the parameter (weight matrix) variance–covariance matrix and calculating parameter standard errors. The simulation study indicated that REML estimation is important for a precise estimate of the uncer-

tainty of the reconciled forecasts. Additionally, including the parameter uncertainty in the forecast variance–covariance matrix gave some improvement in forecast skill. Orthogonality and the effect of shrinkage were illustrated along with parameter uncertainty and correlation for high-dimensional parameter vectors in the case study on electricity load forecasting in Sweden.

The presented framework provided a clear path for model reduction. In the case study, the results indicated that many weights could be set to zero. It should be investigated how this would affect the overall accuracy of the reconciled forecasts. How such tests should be conducted will be the subject of future research. In particular, some guidelines on the order of testing should be developed. The impact of shrinkage on the statistical tests also needs investigation.

In this article, we have focused on showing the equivalence between forecast reconciliation and the LM formulation, while leaving the estimation of the forecast variance–covariance as variants of the observed squared deviation between the reconciled forecasts and observations. Using the presented framework, parameterised versions of the bottom-level reconciled variance–covariance matrix could be estimated using ML or REML estimation.

In summary, we believe that the formulation proposed in this article contributes to the understanding of forecast reconciliation. Specifically, the question of the uncertainty of the estimated reconciliation weights has been answered, at least in the Gaussian case. The proposed LM formulation provides many opportunities for future research.

References

Athanasopoulos, G., R. A. Ahmed, and R. J. Hyndman. “Hierarchical forecasts for Australian domestic tourism.” *International Journal of Forecasting*, vol. 25, no. 1 (2009), pp. 146–166.

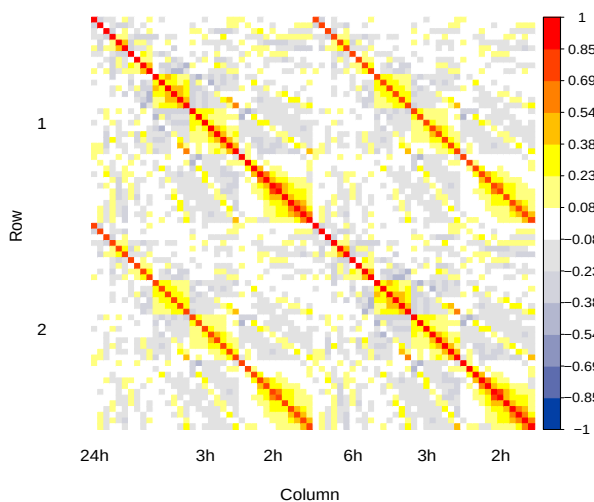


Figure 4: Parameter correlation for the first two rows in the estimated weight matrix.

- Athanasopoulos, G., R. J. Hyndman, N. Kourentzes, and F. Petropoulos. "Forecasting with temporal hierarchies." *European Journal of Operational Research*, vol. 262, no. 1 (2017), pp. 60–74.
- Bjerregård, M., J. Møller, and H. Madsen. "An introduction to multivariate probabilistic forecast evaluation." *Energy and AI*, vol. 4 (2021), p. 100058.
- Di Fonzo, T. and D. Girolimetto. "Forecast combination-based forecast reconciliation: Insights and extensions." *International Journal of Forecasting*, vol. 40, no. 2 (2024), pp. 490–514.
- Eckert, F., R. J. Hyndman, and A. Panagiotelis. "Forecasting Swiss exports using Bayesian forecast reconciliation." *European Journal of Operational Research*, vol. 291, no. 2 (2021), pp. 693–710.
- Gneiting, T. and A. E. Raftery. "Strictly proper scoring rules, prediction, and estimation." *Journal of the American Statistical Association*, vol. 102, no. 477 (2007), pp. 359–378.
- Hollyman, R., F. Petropoulos, and M. E. Tipping. "Understanding forecast reconciliation." *European Journal of Operational Research*, vol. 294, no. 1 (2021), pp. 149–160.
- Hyndman, R. J., R. A. Ahmed, G. Athanasopoulos, and H. L. Shang. "Optimal combination forecasts for hierarchical time series." *Computational Statistics & Data Analysis*, vol. 55, no. 9 (2011), pp. 2579–2589.
- Iranmanesh, A., M. Arashi, and S. M. M. Tabatabaey. "On conditional applications of matrix variate normal distribution." *Iranian Journal of Mathematical Sciences and Informatics*, vol. 5, no. 2 (2010), pp. 33–43.
- Kalivas, J. H. "Overview of two-norm (L2) and one-norm (L1) Tikhonov regularization variants for full wavelength or sparse spectral multivariate calibration models or maintenance." *Journal of Chemometrics*, vol. 26, no. 6 (2012), pp. 218–230.
- Ledoit, O. and M. Wolf. "Improved estimation of the covariance matrix of stock returns with an application to portfolio selection." *Journal of Empirical Finance*, vol. 10, no. 5 (2003), pp. 603–621.
- Madsen, H. *Time Series Analysis*. Chapman and Hall (2008).
- Madsen, H. and P. Thyregod. *Introduction to general and generalized linear models*. CRC Press (2011).
- Møller, J. K., P. Nystrup, and H. Madsen. "Likelihood-based inference in temporal hierarchies." *International Journal of Forecasting*, vol. 40, no. 2 (2024), pp. 515–531.
- Nystrup, P., E. Lindström, J. K. Møller, and H. Madsen. "Dimensionality reduction in forecasting with temporal hierarchies." *International Journal of Forecasting*, vol. 37, no. 3 (2021), pp. 1127–1146.
- Nystrup, P., E. Lindström, P. Pinson, and H. Madsen. "Temporal hierarchies with autocorrelation for load forecasting." *European Journal of Operational Research*, vol. 280, no. 3 (2020), pp. 876–888.
- Panagiotelis, A., G. Athanasopoulos, P. Gamakumara, and R. J. Hyndman. "Forecast reconciliation: A geometric view with new insights on bias correction." *International Journal of Forecasting*, vol. 37, no. 1 (2021), pp. 343–359.
- Petersen, K. B. and M. S. Pedersen. *The Matrix Cookbook*. Technical University of Denmark (2006). URL <http://www2.compute.dtu.dk/pubdb/pubs/3274-full.html>. Version 20051003.
- Pritularga, K. F., I. Svetunkov, and N. Kourentzes. "Stochastic coherency in forecast reconciliation." *International Journal of Production Economics*, vol. 240 (2021), p. 108221.
- Rao, C. *Linear Statistical Inference and its Applications*. John Wiley & Sons (1973). Second edition.
- Schafer, J., R. Opgen-Rhein, V. Zuber, M. Ahdesmaki, A. P. D. Silva, and K. Strimmer. *corpcor: Efficient Estimation of Covariance and (Partial) Correlation* (2021). URL <https://CRAN.R-project.org/package=corpcor>. R package

version 1.6.10.

- Schäfer, J. and K. Strimmer. “A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics.” *Statistical Applications in Genetics and Molecular Biology*, vol. 4, no. 1 (2005), p. 32.
- Scheuerer, M. and T. M. Hamill. “Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities.” *Monthly Weather Review*, vol. 143, no. 4 (2015), pp. 1321–1334.
- van Wieringen, W. N. *Lecture notes on ridge regression* (2023). URL <https://arxiv.org/pdf/1509.09169.pdf>. Version 0.60, June 27, 2023.
- Wickramasuriya, S. L., G. Athanasopoulos, and R. J. Hyndman. “Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization.” *Journal of the American Statistical Association*, vol. 114, no. 526 (2019), pp. 804–819.

A. Nomenclature

Symbol	Explanation
ML	Maximum likelihood (4)
LM	General linear model (4)
REML	Restricted maximum likelihood (4)
MAP	Maximum a posteriori (6)
GLS	Generalized least square (13)
LogS	Log-score (66)
$V s_p$	Variogram score (70)
RRMSE	Relative root mean square error (71)
relVs	Relative variogram score (72)
$n \in \mathbb{N}$	Dimension of the base forecast (1)
$m \in \mathbb{N}$	Dimension of the bottom level observations (1)
$T \in \mathbb{N}$	Number of observations in the training set (5)
$p_i \in \mathbb{N}$	Number of parameters for modeling bottom-level observation i (11)
$p = \sum_{i=1}^m p_i \in \mathbb{N}$	Total number of parameters for modeling bottom-level observations (5)
$\bar{p} = \frac{1}{m} \sum_{i=1}^m p_i \in \mathbb{N}$	Average number of regression parameters (C.22)
$\lambda \in [0, 1]$	Shrinkage parameter (35)
$v \in \mathbb{R}_+$	Degrees of freedom for prior distribution of Σ_r (48)
$q_I \in \mathbb{N}$	Number of element in an index set I (29)
$\epsilon_t^h \in \mathbb{R}^n$	Reconciliation error at time t (1)
$\epsilon_t, \mathbf{e}_t \in \mathbb{R}^m$	Collection of regression errors at time t , (7), and Theorem 2.1
$\epsilon \in \mathbb{R}^{Tm}$	Collection of all regression errors (5)
$\tilde{\mathbf{y}}_t \in \mathbb{R}^m$	Reconciled forecast on the bottom level at time t (1).
$\hat{\mathbf{y}}_t \in \mathbb{R}^n$	Base forecast on all levels at time t (1)
$\hat{\mathbf{y}}_{\text{bot}} = \text{vec}(\hat{\mathbf{Y}}_{\text{bot}}^\top) \in \mathbb{R}^{Tm}$	Collection of all bottom-level base forecasts (5)
$\mathbf{y} = \text{vec}(\mathbf{Y}^\top) \in \mathbb{R}^{m \cdot T}$	Collection of all bottom-level observations (5)

$\mathbf{y}_t \in \mathbb{R}^m$	Bottom-level observation at time t (5)
$\tilde{\mathbf{y}}_{F,t} = \mathbf{S}\tilde{\mathbf{y}}_t \in \mathbb{R}^n$	Reconciled forecast on all levels at time t (8).
$\mathbf{y}_{i,\cdot} \in \mathbb{R}^T$	Collection of all observations at bottom level i (D.8)
$\hat{\mathbf{y}}_{\text{bot},t}$	Bottom-level base forecast at time t (21).
$\hat{\mathbf{y}}_{\text{top},t}$	Top-level (i.e., excluding bottom-level) base forecast at time t (21).
$\mathbf{x}_{i,t} \in \mathbb{R}^{p_i}$	One row in the design matrix $\mathbf{X}_{i,\cdot}$ (11)
$\boldsymbol{\beta}, \hat{\boldsymbol{\beta}} \in \mathbb{R}^p$	Coefficients in the regression model (5), (8)
$\hat{\boldsymbol{\beta}}_i \in \mathbb{R}^{p_i}$	Coefficients in the regression model for bottom-level observation i (18)
$\boldsymbol{\beta}_{\text{top}}, \hat{\boldsymbol{\beta}}_{\text{top}} \in \mathbb{R}^{m(n-m)}$	Coefficients in the regression model corresponding to top-level weights (21)
$\boldsymbol{\beta}_{0,\text{top}}$	Prior mean for $\boldsymbol{\beta}_{\text{top}}$ (39).
$\boldsymbol{\beta}^m = \text{vec}^{-1}(\boldsymbol{\beta})$	Matrix version of the regression parameters (may appear with a subscript) (18)
$\mathbf{Y} = \text{vec}^{-1}(\mathbf{y}) \in \mathbb{R}^{T \times m}$	Collection of all bottom-level observations (18)
$\hat{\mathbf{Y}} = [\hat{\mathbf{Y}}_{\text{top}} \quad \hat{\mathbf{Y}}_{\text{bot}}] \in \mathbb{R}^{T \times n}$	Collection of all base forecasts (Corollary 2.1)
$\tilde{\mathbf{Y}} \in \mathbb{R}^{T \times m}$	Collection of all bottom-level reconciled forecast (27a)
$\hat{\mathbf{Y}}_{\text{bot}} = \text{vec}^{-1}(\hat{\mathbf{y}}_{\text{bot}}) \in \mathbb{R}^{T \times m}$	Collection of all bottom-level base forecasts (40)
$\hat{\mathbf{Y}}_I \in \mathbb{R}^{T \times q_I}$	Column selection of $\hat{\mathbf{Y}}$, Lemma 2.1
$\hat{\mathbf{Y}}_{\text{top}} \in \mathbb{R}^{T \times (n-m)}$	Collection of top level base forecasts (i.e. excluding the bottom levels) (40).
$\mathbf{SSE}, \mathbf{SS}_{\text{mod}}$	Sum of square matrices (61)
$\mathbf{S} = [\mathbf{S}_{\text{top}}^\top \quad \mathbf{I}]^\top \in \mathbb{R}^{n \times m}$	Summation matrix (1)
$\mathbf{S}_{\text{top}} \in \mathbb{R}^{(n-m) \times m}$	Top-level summation matrix (i.e. excluding the bottom level) (2)
$\mathbf{S}_I \in \mathbb{R}^{q_I \times m}$	A row selection of $\mathbf{S}$ defined by the indexset I , Lemma 2.1
$\mathbf{X} \in \mathbb{R}^{Tm \times p}$	Design matrix (5)
$\mathbf{X}_{\cdot,t} \in \mathbb{R}^{m \times p}$	Design matrix at time t (7)
$\mathbf{X}_{i,\cdot} \in \mathbb{R}^{T \times p_i}$	Design matrix for bottom observation i (12)

$P = [P_{\text{top}} \ P_{\text{bot}}] \in \mathbb{R}^{m \times n}$	The weight matrix s.t. $\tilde{Y} = P\hat{Y}$ (3)
$P(\lambda)$	Weight matrix based on shrinkage using λ (35)
$H \in \mathbb{R}^{T \times T}$	Projection matrix, Lemma (2.1)
$\Sigma_h \in \mathbb{R}^{n \times n}$	Variance–covariance of the reconciliation error (1)
$\hat{\Sigma}_h \in \mathbb{R}^{n \times n}$	The specific form given in (20)
$\Sigma_r \in \mathbb{R}^{m \times m}$	Variance–covariance of the regression error (at time t) (7)
σ^{ij}	Element (i, j) of Σ_r^{-1} (C.4)
$\hat{\Sigma}_{r,\text{ML}}, \hat{\Sigma}_{r,\text{REML}} \in \mathbb{R}^{m \times m}$	ML and REML estimates/estimators for Σ_r (14) and (17)
$\Sigma \in \mathbb{R}^{T \cdot m \times T \cdot m}$	Variance–covariance of the regression error (5)
$\hat{\Sigma}_s = (1 - \lambda)\hat{\Sigma}_h + \lambda\hat{\Sigma}_h^d \in \mathbb{R}^{n \times n}$	Variance–covariance of the base-forecast error (35)
$\hat{\Sigma}_{h,\text{bot}}^d \in \mathbb{R}^{m \times m}$	Σ_h^d for bottom level (39)
$\hat{\Sigma}_{h,\text{top}}^d \in \mathbb{R}^{(n-m) \times (n-m)}$	Σ_h^d for top level (39)
$\hat{\Sigma}_{r,\text{MAP}}, \hat{\Sigma}_{r,\text{shrink}}, \hat{\Sigma}_{r,s\text{REML}}, \hat{\Sigma}_{r,\text{par}}$	Versions of the MAP estimator for Σ_r , (43), (49), (51), (56)
$\Sigma_\beta = \Sigma_r \otimes \Sigma_{\beta,0}$	Prior variance for β_T (I.9)
$\Sigma_{\beta,0} \in \mathbb{R}^{(n-m) \times (n-m)}$	Related to prior variance for β_T (39)
Ψ	Prior parameter for Σ_r (48)
$W_{1,\lambda} = \left(X^\top \Sigma^{-1} X + \Sigma_\beta^{-1} \right)^{-1}$	Used to derive the variance of MAP estimator (57)
$W_2 = X^\top \Sigma^{-1} X$	Used to derive the variance of MAP estimator (57)

Table A.3: Symbols used in the article. The first usage is marked in parenthesis. If first usage is not in an equation the appearance will be around the referred equation.

B. Some useful relations

B.1. Some results from matrix algebra

As the proofs in this appendix depend heavily on properties of Kronecker products and vectorisation, we have collected the most important ones here. All expression are collected from Petersen and Pedersen (2006), where many useful relations used in this article can be found.

The vectorisation of a matrix $\mathbf{P} \in \mathbb{R}^{m \times n}$ is given by the invertible map $\mathbb{R}^{m \times n} \rightarrow \mathbb{R}^{mn}$

$$\text{vec}(\mathbf{P}) = [P_{11}, \dots, P_{m1}, P_{12}, \dots, P_{m2}, \dots, P_{1n}, \dots, P_{mn}]^\top. \quad (\text{B.1})$$

The inverse vectorisation of the vector $\mathbf{y} \in \mathbb{R}^{mn}$ is given by the invertible map $\mathbb{R}^{mn} \rightarrow \mathbb{R}^{m \times n}$

$$\text{vec}_{m \times n}^{-1}(\mathbf{y}) = \begin{bmatrix} y_1 & y_{m+1} & \cdots & y_{n \cdot (m-1) + 1} \\ \vdots & \vdots & & \vdots \\ y_m & y_{2m} & \cdots & y_{n \cdot m} \end{bmatrix}. \quad (\text{B.2})$$

As the target space ($\mathbb{R}^{m \times n}$) should be clear from the context we will omit the subscript and use the simpler notation $\text{vec}^{-1}(\cdot)$.

The vectorisation of a product of matrices $\mathbf{A} \in \mathbb{R}^{k \times l}$, $\mathbf{B} \in \mathbb{R}^{l \times m}$, and $\mathbf{C} \in \mathbb{R}^{m \times n}$ can be written as

$$\text{vec}(\mathbf{ABC}) = (\mathbf{C}^\top \otimes \mathbf{A}) \text{vec}(\mathbf{B}) \quad (\text{B.3})$$

$$= (\mathbf{I}_n \otimes (\mathbf{AB})) \text{vec}(\mathbf{C}) \quad (\text{B.4})$$

$$= ((\mathbf{C}^\top \mathbf{B}^\top) \otimes \mathbf{I}_k) \text{vec}(\mathbf{A}) \quad (\text{B.5})$$

$$\text{vec}(\mathbf{AB}) = (\mathbf{I}_m \otimes \mathbf{A}) \text{vec}(\mathbf{B}) \quad (\text{B.6})$$

$$= (\mathbf{B} \otimes \mathbf{I}_k) \text{vec}(\mathbf{A}). \quad (\text{B.7})$$

Further, the following relations apply to matrices of appropriate dimensions and invertiability

$$(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1} \quad (\text{B.8})$$

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC}) \otimes (\mathbf{BD}). \quad (\text{B.9})$$

If $\mathbf{A} \in \mathbb{R}^{k \times k}$ and $\mathbf{B} \in \mathbb{R}^{l \times l}$ then

$$|\mathbf{A} \otimes \mathbf{B}| = |\mathbf{A}|^l |\mathbf{B}|^k. \quad (\text{B.10})$$

For derivatives wrt. variance–covariance matrices the following relations hold (Petersen and Pedersen, 2006, (61), (57), (100))

$$\frac{\partial \mathbf{a}^\top \boldsymbol{\Sigma}^{-1} \mathbf{b}}{\partial \boldsymbol{\Sigma}} = -\boldsymbol{\Sigma}^{-1} \mathbf{a} \mathbf{b}^\top \boldsymbol{\Sigma}^{-1} \quad (\text{B.11})$$

$$\frac{\partial \log |\boldsymbol{\Sigma}|}{\partial \boldsymbol{\Sigma}} = \boldsymbol{\Sigma}^{-1} \quad (\text{B.12})$$

$$\frac{\partial \text{Tr}(\boldsymbol{\Psi} \boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}} = \boldsymbol{\Psi}. \quad (\text{B.13})$$

Combining (B.13) and Petersen and Pedersen (2006, (59), (124)) while assuming that Σ and Ψ are symmetric yields

$$\begin{aligned}\frac{\partial \text{Tr}(\Psi \Sigma^{-1})}{\partial \Sigma} &= -\Sigma^{-1} \Psi \Sigma^{-1} \\ \frac{\partial \text{Tr}(\Psi_1 \Sigma^{-1} \Psi_2)}{\partial \Sigma} &= -\Sigma^{-1} \Psi_1^\top \Psi_2^\top \Sigma^{-1}.\end{aligned}\tag{B.14}$$

For products of vectors the following (Petersen and Pedersen, 2006, eq. (521)) holds

$$\text{vec}(\mathbf{A})^\top \text{vec}(\mathbf{B}) = \text{Tr}(\mathbf{A}^\top \mathbf{B}).\tag{B.15}$$

The following relation, which is related to the matrix normal distribution (Iranmanesh et al., 2010), will also be useful. Using (B.3) and (B.15), we can write

$$\begin{aligned}\mathbf{y}^\top (\Sigma^{-1} \otimes \Omega^{-1}) \mathbf{y} &= \mathbf{y}^\top \text{vec}(\Omega^{-1} \text{vec}^{-1}(\mathbf{y}) \Sigma^{-1}) \\ &= \text{Tr} \left(\text{vec}^{-1}(\mathbf{y})^\top \Omega^{-1} \text{vec}^{-1}(\mathbf{y}) \Sigma^{-1} \right) \\ &= \text{Tr} \left(\mathbf{Y}^\top \Omega^{-1} \mathbf{Y} \Sigma^{-1} \right).\end{aligned}\tag{B.16}$$

B.2. Probability densities, log-likelihood functions and log-priors

In this work we use the following multivariate distributions: the multivariate normal (Gaussian), the inverse Wishart, and the Wishart distribution. Only the first two are used for likelihood/MAP estimation, but for completeness we include all three in the discussion below.

B.2.1. Multivariate normal distribution

If a random variable ($\mathbf{y} \in \mathbb{R}^n$) follows a multivariate normal distribution with mean $\boldsymbol{\mu}$ and variance-covariance Σ , we write $\mathbf{y} \sim N(\boldsymbol{\mu}, \Sigma)$, with $\boldsymbol{\mu} \in \mathbb{R}^n$, $\Sigma \in \mathbb{R}^{n \times n}$, and Σ positive semi-definite. The probability density function (pdf.), $\phi(\cdot)$, of the multivariate normal is defined when Σ is positive definite and is given by

$$\begin{aligned}\phi(\mathbf{y}; \boldsymbol{\mu}, \Sigma) &= \frac{1}{(2\pi)^{n/2} \sqrt{|\Sigma|}} e^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{y}-\boldsymbol{\mu})} \\ &= \frac{1}{(2\pi)^{n/2} \sqrt{|\Sigma|}} e^{-\frac{1}{2}\text{Tr}((\mathbf{y}-\boldsymbol{\mu})\Sigma^{-1}(\mathbf{y}-\boldsymbol{\mu})^\top)}.\end{aligned}\tag{B.17}$$

For likelihood estimation we use parameterised versions of the mean and variance, i.e., $\boldsymbol{\mu} := \boldsymbol{\mu}(\boldsymbol{\beta})$ and $\boldsymbol{\Sigma} := \boldsymbol{\Sigma}(\boldsymbol{\psi})$. The log-likelihood is

$$\begin{aligned} l(\boldsymbol{\beta}, \boldsymbol{\psi}) &= \log(\phi(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})) \\ &\propto -\frac{1}{2} \log(|\boldsymbol{\Sigma}(\boldsymbol{\psi})|) - \frac{1}{2}(\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta}))^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\psi})(\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta})). \end{aligned} \quad (\text{B.18})$$

For fixed $\boldsymbol{\psi}$ or $\boldsymbol{\beta}$ we will use the notation

$$\begin{aligned} l(\boldsymbol{\beta}) &\propto -\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta}))^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\psi})(\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta})) \\ l(\boldsymbol{\psi}) &\propto -\frac{1}{2} \log(|\boldsymbol{\Sigma}(\boldsymbol{\psi})|) - \frac{1}{2}(\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta}))^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\psi})(\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta})). \end{aligned} \quad (\text{B.19})$$

When used as a prior distribution, i.e., $\mathbf{y}|\boldsymbol{\beta} \sim N(\boldsymbol{\mu}(\boldsymbol{\beta}), \boldsymbol{\Sigma}(\boldsymbol{\psi}))$, $\boldsymbol{\beta} \sim N(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta(\boldsymbol{\psi}))$, we will use the notation

$$l_{MAP}(\boldsymbol{\beta}, \boldsymbol{\psi}) = \log(\phi(\mathbf{y}; \boldsymbol{\mu}(\boldsymbol{\beta}), \boldsymbol{\Sigma}(\boldsymbol{\psi}))) + \log(\phi(\boldsymbol{\beta}; \boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta(\boldsymbol{\psi}))). \quad (\text{B.20})$$

B.2.2. The inverse Wishart distribution

If a random matrix $\boldsymbol{\Sigma}_r \in \mathbb{R}^{m \times m}$ ($\boldsymbol{\Sigma}_r$ positive definite) follows an inverse Wishart distribution with scale matrix $\boldsymbol{\Psi} > 0$ and degrees of freedom $v > m - 1$, we write $\boldsymbol{\Sigma}_r \sim \mathcal{W}^{-1}(\boldsymbol{\Psi}, v)$ and the density is given by

$$f_{\boldsymbol{\Sigma}_r}(\boldsymbol{\Sigma}_r) = \frac{|\boldsymbol{\Psi}|^{v/2}}{2^{vm/2} \Gamma_m\left(\frac{v}{2}\right)} |\boldsymbol{\Sigma}_r|^{-\frac{m+v+1}{2}} e^{-\frac{1}{2} \text{Tr}(\boldsymbol{\Psi} \boldsymbol{\Sigma}_r^{-1})}, \quad (\text{B.21})$$

where $\Gamma_m(\cdot)$ is the multivariate Gamma function. The inverse Wishart is the conjugate prior for the variance–covariance matrix in the multivariate normal. When used as a prior distribution (i.e., $\boldsymbol{\Psi}$ and v fixed) we will use the form

$$\log(f_{\boldsymbol{\Sigma}_r}(\boldsymbol{\Sigma}_r)) \propto -\frac{m+v+1}{2} \log(|\boldsymbol{\Sigma}_r|) - \frac{1}{2} \text{Tr}(\boldsymbol{\Psi} \boldsymbol{\Sigma}_r^{-1}). \quad (\text{B.22})$$

In this work we use the inverse Wishart as a prior in the setting $\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\Sigma}_r \sim N(\boldsymbol{\mu}(\boldsymbol{\beta}), \boldsymbol{\Sigma}(\boldsymbol{\Sigma}_r))$, $\boldsymbol{\beta}|\boldsymbol{\Sigma}_r \sim N(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta(\boldsymbol{\Sigma}_r))$, $\boldsymbol{\Sigma}_r \sim \mathcal{W}^{-1}(\boldsymbol{\Psi}, v)$, and we use the notation

$$l_{MAP}(\boldsymbol{\beta}, \boldsymbol{\Sigma}_r) = \log(\phi(\mathbf{y}; \boldsymbol{\mu}(\boldsymbol{\beta}), \boldsymbol{\Sigma}(\boldsymbol{\Sigma}_r))) + \log(\phi(\boldsymbol{\beta}; \boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta(\boldsymbol{\Sigma}_r))) + \log(f_{\boldsymbol{\Sigma}_r}(\boldsymbol{\Sigma}_r)). \quad (\text{B.23})$$

B.2.3. The Wishart distribution

In general we use the sum of squares definition of the Wishart distribution. The following definition and lemma is adapted from Rao (1973). We start with the general definition of the Wishart distribution.

Definition B.1 (Wishart distribution). Let the matrix $\mathbf{Z} = [\mathbf{z}_1 \ \dots \ \mathbf{z}_T]^\top \in \mathbb{R}^{T \times p}$ be a collection of normal random variables s.t. $\mathbf{z}_i \sim N(\mathbf{0}, \Sigma_z)$ and iid., then $\mathbf{Q} = \mathbf{Z}^\top \mathbf{Z}$ follows a Wishart distribution with T degrees of freedom and scale parameter Σ_z , $\mathbf{Q} \sim W_p(\Sigma_z, T)$. If $T > p - 1$ and Σ_z is positive definite, then the probability density function exists and $E[\mathbf{Q}] = T\Sigma_z$.

The following lemma relates projections and the Wishart distribution.

Lemma B.1 (Some properties of the Wishart distribution). Let $\mathbf{Z}$ be as in Definition B.1 and $\mathbf{H}$ a projection matrix with $r = \text{Tr}(\mathbf{H})$, then a necessary and sufficient condition for $\mathbf{Z}^\top \mathbf{H} \mathbf{Z}$ to follow a Wishart distribution is that $\frac{\mathbf{v}^\top \mathbf{Z}^\top \mathbf{H} \mathbf{Z} \mathbf{v}}{\sigma_v^2} \sim \chi^2(r)$, for any fixed vector $\mathbf{v}$, and in that case $\mathbf{Z}^\top \mathbf{H} \mathbf{Z} \sim W_p(\Sigma_z, r)$.

Proof. See Rao (1973, 8b) ■

The general definition of the Wishart distribution used in this article is given in Definition B.1. For completeness we also state the pdf. If $\mathbf{Q} \in \mathbb{R}^{m \times m}$ ($\mathbf{Q}$ positive definite) follows a Wishart distribution with positive definite scale matrix Ψ and degrees of freedom v , we write $\mathbf{Q} \sim \mathcal{W}(\Psi, v)$. The pdf is defined for $v > m - 1$ and is given by

$$f_{\mathbf{Q}}(\mathbf{Q}) = \frac{|\Psi|^{-v/2}}{2^{vm/2} \Gamma_m\left(\frac{v}{2}\right)} |\mathbf{Q}|^{\frac{m+v-1}{2}} e^{-\frac{1}{2} \text{Tr}(\Psi \mathbf{Q})}. \quad (\text{B.24})$$

C. Proof of Theorem 2.1

The log-likelihood is given by

$$\begin{aligned} l(\Sigma_r; \mathbf{e}) &\propto -\frac{1}{2} \log(|\mathbf{I}_T \otimes \Sigma_r|) - \frac{1}{2} \mathbf{e}^\top (\mathbf{I}_T \otimes \Sigma_r^{-1}) \mathbf{e} \\ &= -\frac{1}{2} \log(|\Sigma_r|^T) - \frac{1}{2} \sum_{t=1}^T \mathbf{e}_t^\top \Sigma_r^{-1} \mathbf{e}_t, \end{aligned} \quad (\text{C.1})$$

with

$$\mathbf{e}_t = \mathbf{y}_t - \mathbf{X}_{\cdot,t} \hat{\boldsymbol{\beta}}. \quad (\text{C.2})$$

$\hat{\boldsymbol{\beta}}$ is the ML (or weighted least square) estimate of $\boldsymbol{\beta}$ is well-known and, given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top (\mathbf{I}_T \otimes \boldsymbol{\Sigma}_r^{-1}) \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{I}_T \otimes \boldsymbol{\Sigma}_r^{-1}) \mathbf{y}. \quad (\text{C.3})$$

By direct matrix multiplications, using the diagonal like structure in (11), we find the matrix $\mathbf{X}^\top \boldsymbol{\Sigma}^{-1} \mathbf{X}$ as

$$\begin{aligned} \mathbf{X}^\top \boldsymbol{\Sigma}^{-1} \mathbf{X} &= \begin{bmatrix} \sigma^{11} \sum_{t=1}^T \mathbf{x}_{1,t} \mathbf{x}_{1,t}^\top & \sigma^{12} \sum_{t=1}^T \mathbf{x}_{1,t} \mathbf{x}_{2,t}^\top & \cdots & \sigma^{1m} \sum_{t=1}^T \mathbf{x}_{1,t} \mathbf{x}_{m,t}^\top \\ \sigma^{12} \sum_{t=1}^T \mathbf{x}_{2,t} \mathbf{x}_{1,t}^\top & \sigma^{22} \sum_{t=1}^T \mathbf{x}_{2,t} \mathbf{x}_{2,t}^\top & \cdots & \sigma^{2m} \sum_{t=1}^T \mathbf{x}_{2,t} \mathbf{x}_{m,t}^\top \\ \vdots & & \ddots & \\ \sigma^{1m} \sum_{t=1}^T \mathbf{x}_{m,t} \mathbf{x}_{1,t}^\top & \cdots & & \sigma^{mm} \sum_{t=1}^T \mathbf{x}_{m,t} \mathbf{x}_{m,t}^\top \end{bmatrix} \\ &= \begin{bmatrix} \sigma^{11} \mathbf{X}_{1,\cdot}^\top \mathbf{X}_{1,\cdot} & \sigma^{12} \mathbf{X}_{1,\cdot}^\top \mathbf{X}_{2,\cdot} & \cdots & \sigma^{1m} \mathbf{X}_{1,\cdot}^\top \mathbf{X}_{m,\cdot} \\ \sigma^{12} \mathbf{X}_{2,\cdot}^\top \mathbf{X}_{1,\cdot} & \sigma^{22} \mathbf{X}_{2,\cdot}^\top \mathbf{X}_{2,\cdot} & \cdots & \sigma^{2m} \mathbf{X}_{2,\cdot}^\top \mathbf{X}_{m,\cdot} \\ \vdots & & \ddots & \\ \sigma^{1m} \mathbf{X}_{m,\cdot}^\top \mathbf{X}_{1,\cdot} & \cdots & & \sigma^{mm} \mathbf{X}_{m,\cdot}^\top \mathbf{X}_{m,\cdot} \end{bmatrix}, \end{aligned} \quad (\text{C.4})$$

where $\sigma^{ij} = (\boldsymbol{\Sigma}^{-1})_{ij}$.

Using (B.3) we get

$$(\mathbf{I}_T \otimes \boldsymbol{\Sigma}_r^{-1}) \mathbf{y} = \text{vec}(\boldsymbol{\Sigma}_r^{-1} \mathbf{Y}^\top). \quad (\text{C.5})$$

We treat the special case below.

Special case 1: $\mathbf{x}_{i,t} = \mathbf{x}_{j,t} = \mathbf{x}_t, \forall (i, j)$. In this case we get (using (C.4))

$$\mathbf{X}^\top \boldsymbol{\Sigma}^{-1} \mathbf{X} = \boldsymbol{\Sigma}_r^{-1} \otimes \sum_{i=1}^T \mathbf{x}_t \mathbf{x}_t^\top = \boldsymbol{\Sigma}_r^{-1} \otimes (\mathbf{X}_{1,\cdot}^\top \mathbf{X}_{1,\cdot}), \quad (\text{C.6})$$

and therefore

$$(\mathbf{X}^\top \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} = \boldsymbol{\Sigma}_r \otimes \left(\sum_{i=1}^T \mathbf{x}_t \mathbf{x}_t^\top \right)^{-1} = \boldsymbol{\Sigma}_r \otimes (\mathbf{X}_{1,\cdot}^\top \mathbf{X}_{1,\cdot})^{-1}. \quad (\text{C.7})$$

Setting $\boldsymbol{\xi} = \boldsymbol{\Sigma}_r^{-1} \mathbf{Y}^\top$ we can write

$$\begin{aligned} \mathbf{X}^\top \text{vec}(\boldsymbol{\xi}) &= \begin{bmatrix} \mathbf{x}_1 & \mathbf{0} & \mathbf{0} & \mathbf{x}_2 & \cdots & \mathbf{x}_T & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} & & \ddots & & \ddots \\ \mathbf{0} & \mathbf{0} & \mathbf{x}_1 & \mathbf{0} & \mathbf{x}_2 & \cdots & \mathbf{x}_T \end{bmatrix} \begin{bmatrix} \xi_{11} \\ \vdots \\ \xi_{1m} \\ \xi_{21} \\ \vdots \\ \xi_{Tm} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{x}_1 \xi_{11} + \mathbf{x}_2 \xi_{21} + \cdots \mathbf{x}_T \xi_{T1} \\ \vdots \\ \mathbf{x}_1 \xi_{1m} + \mathbf{x}_2 \xi_{2m} + \cdots \mathbf{x}_T \xi_{Tm} \end{bmatrix}. \end{aligned} \quad (\text{C.8})$$

Hence, the order can be rearranged to

$$\begin{aligned} \mathbf{X}^\top \text{vec}(\boldsymbol{\xi}) &= \begin{bmatrix} \mathbf{X}_{1,\cdot}^\top & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_{1,\cdot}^\top & \cdots & \mathbf{0} \\ \vdots & & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & & \mathbf{X}_{1,\cdot}^\top \end{bmatrix} \begin{bmatrix} \xi_{11} \\ \vdots \\ \xi_{T1} \\ \xi_{12} \\ \vdots \\ \xi_{Tm} \end{bmatrix} \\ &= (\mathbf{I}_m \otimes \mathbf{X}_{1,\cdot}^\top) \text{vec}(\boldsymbol{\xi}^\top) \\ &= (\mathbf{I}_m \otimes \mathbf{X}_{1,\cdot}^\top) \text{vec}(\mathbf{Y} \boldsymbol{\Sigma}_r^{-1}) \\ &= \text{vec}(\mathbf{X}_{1,\cdot}^\top \mathbf{Y} \boldsymbol{\Sigma}_r^{-1}), \end{aligned} \quad (\text{C.9})$$

and, consequently, (again using (B.3))

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \left[\boldsymbol{\Sigma}_r \otimes (\mathbf{X}_{1,\cdot}^\top \mathbf{X}_{1,\cdot})^{-1} \right] \mathbf{X}^\top (\mathbf{I}_T \otimes \boldsymbol{\Sigma}_r^{-1}) \mathbf{y} \\ &= \text{vec} \left((\mathbf{X}_{1,\cdot}^\top \mathbf{X}_{1,\cdot})^{-1} \mathbf{X}_{1,\cdot}^\top \mathbf{Y} \right). \end{aligned} \quad (\text{C.10})$$

This proves that $\hat{\boldsymbol{\beta}}$ is independent of $\boldsymbol{\Sigma}_r$.

Maximum likelihood estimate of Σ_r

The log-likelihood wrt. Σ_r is given by

$$\begin{aligned} l(\Sigma_r; e) &\propto -\frac{1}{2} \log(|I_T \otimes \Sigma_r|) - \frac{1}{2} e^\top (I_T \otimes \Sigma_r^{-1}) e \\ &= -\frac{1}{2} \log(|\Sigma_r|^T) - \frac{1}{2} \sum_{t=1}^T e_t^\top \Sigma_r^{-1} e_t. \end{aligned} \quad (\text{C.11})$$

In the general case, e_t is a function of Σ_r ; but as shown above, it is independent of Σ_r in the special Case 1.

The derivative wrt. Σ_r of the log-likelihood is

$$\frac{\partial l}{\partial \Sigma_r} = -\frac{T}{2} \Sigma_r^{-1} + \frac{1}{2} \Sigma_r^{-1} \sum_{t=1}^T e_t e_t^\top \Sigma_r^{-1}, \quad (\text{C.12})$$

and, hence, for fixed β , the maximum likelihood estimate of Σ_r is

$$\hat{\Sigma}_{ML,r} = \frac{1}{T} \sum_{t=1}^T e_t e_t^\top. \quad (\text{C.13})$$

In the general case where the error e_t is a function of $\hat{\Sigma}_r$, iterations are needed in order to find the best estimate.

REML estimate of Σ_r

The restricted log-likelihood wrt. Σ_r is given by

$$l_{\text{REML}}(\Sigma_r; e) \propto -\frac{1}{2} \log(|\Sigma_r|^T) - \frac{1}{2} \sum_{t=1}^T e_t^\top \Sigma_r^{-1} e_t - \frac{1}{2} \log(|\mathbf{X}^\top \Sigma^{-1} \mathbf{X}^\top|). \quad (\text{C.14})$$

We divide the answer into a special case and the general case:

Special case: If $x_{i,t} = x_{j,t} = x_t \in \mathbb{R}^{\bar{p}}$, then $\hat{\beta}$ and therefore e are independent from Σ_r , which (using (C.6)) leads to

$$\begin{aligned} l_{\text{REML}}(\Sigma_r; e) &\propto -\frac{1}{2} \log(|\Sigma_r|^T) - \frac{1}{2} \sum_{t=1}^T e_t^\top \Sigma_r^{-1} e_t - \frac{1}{2} \log \left| \Sigma_r^{-1} \otimes \left(\sum_{t=1}^T x_t x_t^\top \right) \right| \\ &= -\frac{T}{2} \log(|\Sigma_r|) - \frac{1}{2} \sum_{t=1}^T e_t^\top \Sigma_r^{-1} e_t - \frac{1}{2} \log \left| \Sigma_r^{-1} \right|^{\bar{p}}, \end{aligned} \quad (\text{C.15})$$

and

$$\frac{\partial l_{\text{REML}}}{\partial \Sigma_r} = -\frac{T}{2} \Sigma_r^{-1} + \frac{1}{2} \Sigma_r^{-1} \sum_{t=1}^T \mathbf{e}_t \mathbf{e}_t^\top \Sigma_r^{-1} + \frac{\bar{p}}{2} \Sigma_r^{-1}. \quad (\text{C.16})$$

The resulting REML estimate of Σ_r is

$$\hat{\Sigma}_{r,\text{REML}} = \frac{1}{T - \bar{p}} \sum_{t=1}^T \mathbf{e}_t \mathbf{e}_t^\top. \quad (\text{C.17})$$

General case: In the general case we need the derivative of the REML term

$$\frac{\partial \log |\mathbf{X}^\top \Sigma^{-1} \mathbf{X}|}{\partial \Sigma_r} = \frac{\partial \log |\mathbf{X}^\top \Sigma^{-1} \mathbf{X}|}{\partial \Sigma_r^{-1}} \frac{\partial \Sigma^{-1}}{\partial \Sigma_r}. \quad (\text{C.18})$$

To that end we calculate (using Petersen and Pedersen, 2006, (46))

$$\frac{\partial \log |\mathbf{X}^\top \Sigma^{-1} \mathbf{X}|}{\partial \sigma^{ij}} = \text{Tr} \left((\mathbf{X}^\top \Sigma^{-1} \mathbf{X})^{-1} \frac{\partial \mathbf{X}^\top \Sigma^{-1} \mathbf{X}}{\partial \sigma^{ij}} \right). \quad (\text{C.19})$$

With $\mathbf{X}^\top \Sigma^{-1} \mathbf{X}$ as in (C.4) we get

$$\frac{\partial \mathbf{X}^\top \Sigma^{-1} \mathbf{X}}{\partial \sigma^{ij}} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \mathbf{X}_{i,\cdot}^\top \mathbf{X}_{j,\cdot} & 0 \\ 0 & 0 & 0 \end{bmatrix}. \quad (\text{C.20})$$

Therefore, (C.19) can be written as

$$\frac{\partial \log |\mathbf{X}^\top \Sigma^{-1} \mathbf{X}|}{\partial \sigma^{ij}} = \text{Tr} \left((\mathbf{X}^\top \Sigma^{-1} \mathbf{X})_{I_j, I_i}^{-1} \mathbf{X}_{i,\cdot}^\top \mathbf{X}_{j,\cdot} \right), \quad (\text{C.21})$$

with $I_i = \{P_{i-1} + 1, P_i + 2, \dots, P_i + p_i\}$ and $P_i = \sum_{l=0}^i p_l$, with the convention that $p_0 = 0$. From that we have

$$\begin{aligned} \left(\frac{\partial \log |\mathbf{X}^\top \Sigma^{-1} \mathbf{X}|}{\partial \Sigma_r} \right)_{ij} &= \text{Tr} \left(\frac{\partial \log |\mathbf{X}^\top \Sigma^{-1} \mathbf{X}|}{\partial \Sigma_r^{-1}} \Sigma_r^{-1} \frac{\partial \Sigma_r}{\sigma_{ij}} \Sigma_r^{-1} \right) \\ &= \text{Tr} \left(\Sigma_r^{-1} \frac{\partial \log |\mathbf{X}^\top \Sigma_r^{-1} \mathbf{X}|}{\partial \Sigma_r^{-1}} \Sigma_r^{-1} \frac{\partial \Sigma_r}{\sigma_{ij}} \right). \end{aligned} \quad (\text{C.22})$$

Since $\left(\frac{\partial \Sigma_r}{\sigma_{ij}} \right)_{kl} = 1$ for $k = i, l = j$ and zero otherwise we have

$$\frac{\partial \log |\mathbf{X}^\top \Sigma^{-1} \mathbf{X}|}{\partial \Sigma_r} = \Sigma_r^{-1} \frac{\partial \log |\mathbf{X}^\top \Sigma_r^{-1} \mathbf{X}|}{\partial \Sigma_r^{-1}} \Sigma_r^{-1}, \quad (\text{C.23})$$

and the REML estimate of Σ_r is given as the solution to

$$-\frac{T}{2} \Sigma_r^{-1} + \frac{1}{2} \Sigma_r^{-1} \sum_{t=1}^T \mathbf{e}_t \mathbf{e}_t^\top \Sigma_r^{-1} + \frac{1}{2} \frac{\partial \log |\mathbf{X}^\top \Sigma_r^{-1} \mathbf{X}|}{\partial \Sigma_r} = 0, \quad (\text{C.24})$$

or

$$\begin{aligned}\hat{\Sigma}_{\text{REML},r}^{-1} &= \frac{1}{T} \left(\Sigma_r^{-1} \sum_{t=1}^T \mathbf{e}_t \mathbf{e}_t^\top \Sigma_r^{-1} + \frac{\partial \log |\mathbf{X}^\top \Sigma_r^{-1} \mathbf{X}|}{\partial \Sigma_r} \right) \bigg|_{\Sigma_r = \hat{\Sigma}_{r,\text{REML}}} \\ &= \frac{1}{T} \left(\Sigma_r^{-1} \left(\sum_{t=1}^T \mathbf{e}_t \mathbf{e}_t^\top + \frac{\partial \log |\mathbf{X}^\top \Sigma_r \mathbf{X}|}{\partial \Sigma_r} \right) \Sigma_r^{-1} \right) \bigg|_{\Sigma_r = \hat{\Sigma}_{r,\text{REML}}}.\end{aligned}\quad (\text{C.25})$$

The solution can be written in terms of Σ_r as

$$\hat{\Sigma}_{\text{REML},r} = \frac{1}{T} \left(\sum_{t=1}^T \mathbf{e}_t \mathbf{e}_t^\top + \frac{\partial \log |\mathbf{X}^\top \Sigma_r^{-1} \mathbf{X}|}{\partial \Sigma_r^{-1}} \right) \bigg|_{\Sigma_r = \hat{\Sigma}_{r,\text{REML}}}.\quad (\text{C.26})$$

This is (17) and concludes the proof of Theorem 2.1 and Eq. (17). $\blacksquare$

D. Proof of Theorem 2.2

Consider the regression problem

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}; \quad \boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \mathbf{I}_T \otimes \Sigma_r), \quad (\text{D.1})$$

where $\mathbf{X}_{:,t} = \mathbf{I}_m \otimes \hat{\mathbf{y}}_t^\top$, with the linear constraints $\mathbf{S}^\top \text{vec}^{-1}(\boldsymbol{\beta}) = \mathbf{I}_m$. We use the following notation

$$\hat{\mathbf{Y}} = \begin{bmatrix} \hat{\mathbf{Y}}_{\text{top}} & \hat{\mathbf{Y}}_{\text{bot}} \end{bmatrix}; \quad \mathbf{S} = \begin{bmatrix} \mathbf{S}_{\text{top}} \\ \mathbf{I}_m \end{bmatrix}; \quad \boldsymbol{\beta}^m = \text{vec}^{-1}(\boldsymbol{\beta}) = \begin{bmatrix} \boldsymbol{\beta}_{\text{top}}^m \\ \boldsymbol{\beta}_{\text{bot}}^m \end{bmatrix}.\quad (\text{D.2})$$

The linear constraints can be written as

$$\mathbf{S}^\top \boldsymbol{\beta}^m = \begin{bmatrix} \mathbf{S}_{\text{top}}^\top & \mathbf{I}_m \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_{\text{top}}^m \\ \boldsymbol{\beta}_{\text{bot}}^m \end{bmatrix} = \mathbf{S}_{\text{top}}^\top \boldsymbol{\beta}_{\text{top}}^m + \boldsymbol{\beta}_{\text{bot}}^m = \mathbf{I}_m, \quad (\text{D.3})$$

i.e.,

$$\boldsymbol{\beta}_{\text{bot}}^m = \mathbf{I}_m - \mathbf{S}_{\text{top}}^\top \boldsymbol{\beta}_{\text{top}}^m. \quad (\text{D.4})$$

Using the model definition (D.1) we get

$$\begin{aligned}\mathbf{y}_t &= (\mathbf{I}_m \otimes \hat{\mathbf{y}}_t^\top) \boldsymbol{\beta} + \boldsymbol{\epsilon}_t \\ &= (\mathbf{I}_m \otimes \hat{\mathbf{y}}_t^\top) \text{vec} \left(\begin{bmatrix} \boldsymbol{\beta}_{\text{top}}^m \\ \boldsymbol{\beta}_{\text{bot}}^m \end{bmatrix} \right) + \boldsymbol{\epsilon}_t \\ &= (\mathbf{I}_m \otimes \hat{\mathbf{y}}_t^\top) \text{vec} \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{I}_m \end{bmatrix} + \begin{bmatrix} \mathbf{I}_{n-m} \\ -\mathbf{S}_{\text{top}}^\top \end{bmatrix} \boldsymbol{\beta}_{\text{top}}^m \right) + \boldsymbol{\epsilon}_t.\end{aligned}\quad (\text{D.5})$$

Using (B.6) we get

$$\begin{aligned} \mathbf{y}_t &= \text{vec} \left(\hat{\mathbf{y}}_t^\top \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{I}_m \end{bmatrix} + \begin{bmatrix} \mathbf{I}_{n-m} \\ -\mathbf{S}_{\text{top}}^\top \end{bmatrix} \beta_{\text{top}}^m \right) \right) + \boldsymbol{\epsilon}_t \\ &= \text{vec}(\hat{\mathbf{y}}_{\text{bot},t}^\top) + \text{vec} \left((\hat{\mathbf{y}}_{\text{top},t}^\top - \hat{\mathbf{y}}_{\text{bot},t}^\top \mathbf{S}_{\text{top}}^\top) \beta_{\text{top}}^m \right) + \boldsymbol{\epsilon}_t. \end{aligned} \quad (\text{D.6})$$

Rearranging, we get

$$\begin{aligned} \mathbf{y}_t - \hat{\mathbf{y}}_{\text{bot},t} &= \left(\mathbf{I}_m \otimes (\hat{\mathbf{y}}_{\text{top},t}^\top - \hat{\mathbf{y}}_{\text{bot},t}^\top \mathbf{S}_{\text{top}}^\top) \right) \beta_{\text{top}} + \boldsymbol{\epsilon}_t \\ &= \tilde{\mathbf{X}}_{\cdot,t} \beta_{\text{top}} + \boldsymbol{\epsilon}_t, \end{aligned} \quad (\text{D.7})$$

which is the postulated model. Using (18) the parameter estimates can be written as

$$\begin{aligned} \beta_{\text{top}}^m &= \left[\left(\tilde{\mathbf{X}}_{1,\cdot}^\top \tilde{\mathbf{X}}_{1,\cdot} \right)^{-1} \tilde{\mathbf{X}}_{1,\cdot}^\top \mathbf{y}_{1,\cdot} \quad \dots \quad \left(\tilde{\mathbf{X}}_{1,\cdot}^\top \tilde{\mathbf{X}}_{1,\cdot} \right)^{-1} \tilde{\mathbf{X}}_{1,\cdot}^\top \mathbf{y}_{m,\cdot} \right] \\ &= \left(\tilde{\mathbf{X}}_{1,\cdot}^\top \tilde{\mathbf{X}}_{1,\cdot} \right)^{-1} \tilde{\mathbf{X}}_{1,\cdot}^\top \mathbf{Y}, \end{aligned} \quad (\text{D.8})$$

where $\mathbf{y}_{i,\cdot} \in \mathbb{R}^T$ is the collection of observation at bottom level i at all times.

By inserting $\tilde{\mathbf{X}}_{1,\cdot}$ we get

$$\hat{\beta}_{\text{top}}^m = \left(\left(\hat{\mathbf{Y}}_{\text{top}} - \hat{\mathbf{Y}}_{\text{bot}} \mathbf{S}_{\text{top}}^\top \right)^\top \left(\hat{\mathbf{Y}}_{\text{top}} - \hat{\mathbf{Y}}_{\text{bot}} \mathbf{S}_{\text{top}}^\top \right) \right)^{-1} \left(\hat{\mathbf{Y}}_{\text{top}} - \hat{\mathbf{Y}}_{\text{bot}} \mathbf{S}_{\text{top}}^\top \right)^\top \left(\mathbf{Y} - \hat{\mathbf{Y}}_{\text{bot}} \right). \quad (\text{D.9})$$

The next step is to show that this is equivalent to $\mathbf{P}$. Using Theorem 1 of Wickramasuriya et al. (2019) the usual weight matrix can be written as

$$\begin{aligned} \mathbf{P} &= \left(\mathbf{S}^\top \boldsymbol{\Sigma}_h^{-1} \mathbf{S} \right)^{-1} \mathbf{S}^\top \boldsymbol{\Sigma}_h^{-1} \\ &= \mathbf{J} - \mathbf{J} \boldsymbol{\Sigma}_h \mathbf{U} \left(\mathbf{U}^\top \boldsymbol{\Sigma}_h \mathbf{U} \right)^{-1} \mathbf{U}^\top, \end{aligned} \quad (\text{D.10})$$

where $\mathbf{J} = [\mathbf{0}_{m \times n-m} \quad \mathbf{I}_m]$, and $\mathbf{U} = \begin{bmatrix} \mathbf{I}_{n-m} \\ -\mathbf{S}_{\text{top}}^\top \end{bmatrix}$.

Replacing $\boldsymbol{\Sigma}_h$ by $\hat{\boldsymbol{\Sigma}}_h = \frac{1}{T} \left(\mathbf{Y} \mathbf{S}^\top - \hat{\mathbf{Y}} \right)^\top \left(\mathbf{Y} \mathbf{S}^\top - \hat{\mathbf{Y}} \right)$, we calculate each term of (D.10)

$$\begin{aligned} \mathbf{U}^\top \left(\mathbf{Y} \mathbf{S}^\top - \hat{\mathbf{Y}} \right)^\top &= \begin{bmatrix} \mathbf{I}_{n-m} & -\mathbf{S}_{\text{top}} \end{bmatrix} \left(\mathbf{S} \mathbf{Y}^\top - \hat{\mathbf{Y}}^\top \right) \\ &= \begin{bmatrix} \mathbf{I}_{n-m} & -\mathbf{S}_{\text{top}} \end{bmatrix} \left(\begin{bmatrix} \mathbf{S}_{\text{top}} \\ \mathbf{I}_m \end{bmatrix} \mathbf{Y}^\top - \hat{\mathbf{Y}}^\top \right) \\ &= - \begin{bmatrix} \mathbf{I}_{n-m} & -\mathbf{S}_{\text{top}} \end{bmatrix} \hat{\mathbf{Y}}^\top \\ &= - \left(\hat{\mathbf{Y}}_{\text{top}}^\top - \mathbf{S}_{\text{top}} \hat{\mathbf{Y}}_{\text{bot}}^\top \right). \end{aligned} \quad (\text{D.11})$$

Further,

$$\begin{aligned} J(\mathbf{Y}\mathbf{S}^\top - \hat{\mathbf{Y}})^\top &= [\mathbf{0}_{m \times n-m} \quad \mathbf{I}_m] \left(\begin{bmatrix} \mathbf{S}_{\text{top}} \\ \mathbf{I}_m \end{bmatrix} \mathbf{Y} - \hat{\mathbf{Y}}^\top \right) \\ &= \mathbf{Y}^\top - \hat{\mathbf{Y}}_{\text{bot}}^\top \end{aligned} \quad (\text{D.12})$$

and

$$\begin{aligned} \mathbf{P} &= \mathbf{J} + \left(\mathbf{Y} - \hat{\mathbf{Y}}_{\text{bot}} \right)^\top \left(\hat{\mathbf{Y}}_{\text{top}} - \hat{\mathbf{Y}}_{\text{bot}} \mathbf{S}_{\text{top}}^\top \right) \left(\left(\hat{\mathbf{Y}}_{\text{top}} - \hat{\mathbf{Y}}_{\text{bot}} \mathbf{S}_{\text{top}}^\top \right)^\top \left(\hat{\mathbf{Y}}_{\text{top}} - \hat{\mathbf{Y}}_{\text{bot}} \mathbf{S}_{\text{top}}^\top \right) \right)^{-1} \mathbf{U}^\top \\ &= \mathbf{J} + \left(\hat{\boldsymbol{\beta}}_{\text{top}}^m \right)^\top \mathbf{U}^\top. \end{aligned} \quad (\text{D.13})$$

Since

$$\left(\hat{\boldsymbol{\beta}}^m \right)^\top = \mathbf{J} + \left(\hat{\boldsymbol{\beta}}_{\text{top}}^m \right)^\top \mathbf{U}^\top, \quad (\text{D.14})$$

we can conclude that if $\boldsymbol{\Sigma}_h = \frac{1}{T} \left(\mathbf{Y}\mathbf{S}^\top - \hat{\mathbf{Y}} \right)^\top \left(\mathbf{Y}\mathbf{S}^\top - \hat{\mathbf{Y}} \right)$ then

$$\mathbf{P} = \left(\hat{\boldsymbol{\beta}}^m \right)^\top, \quad (\text{D.15})$$

where $\hat{\boldsymbol{\beta}}^m$ is calculated using usual linear regression with linear constraint $(\mathbf{S}^\top \hat{\boldsymbol{\beta}}^m = \mathbf{I})$. This proves the equivalence between (23) and (21). The ML and REML estimators in (19) follow from (14) and (16) in Theorem 2.1. Further that $\hat{\boldsymbol{\beta}}_{\text{top}}$ is independent of $\boldsymbol{\Sigma}_r$ also follow from Theorem 2.1. ■

E. Proof of Corollary 2.1

We need to show that

$$\begin{aligned} \frac{1}{T} \mathbf{P} \left(\mathbf{Y}\mathbf{S}^\top - \hat{\mathbf{Y}} \right)^\top \left(\mathbf{Y}\mathbf{S}^\top - \hat{\mathbf{Y}} \right) \mathbf{P}^\top &= \frac{1}{T} \left(\mathbf{Y} - \hat{\mathbf{Y}}_{\text{bot}} - \left(\mathbf{Y}_{\text{top}} - \hat{\mathbf{Y}}_{\text{bot}} \mathbf{S}_{\text{top}}^\top \right) \hat{\boldsymbol{\beta}}_{\text{top}}^m \right)^\top \times \\ &\quad \left(\mathbf{Y} - \hat{\mathbf{Y}}_{\text{bot}} - \left(\mathbf{Y}_{\text{top}} - \hat{\mathbf{Y}}_{\text{bot}} \mathbf{S}_{\text{top}}^\top \right) \hat{\boldsymbol{\beta}}_{\text{top}}^m \right). \end{aligned} \quad (\text{E.1})$$

It suffices to show that

$$\left(\mathbf{Y}\mathbf{S}^\top - \hat{\mathbf{Y}} \right) \mathbf{P}^\top = \mathbf{Y} - \hat{\mathbf{Y}}_{\text{bot}} - \left(\mathbf{Y}_{\text{top}} - \hat{\mathbf{Y}}_{\text{bot}} \mathbf{S}_{\text{top}}^\top \right) \hat{\boldsymbol{\beta}}_{\text{top}}^m. \quad (\text{E.2})$$

Using Theorem 2.2, the left-hand side can be written as

$$\begin{aligned}
(\mathbf{Y}\mathbf{S}^\top - \hat{\mathbf{Y}})\mathbf{P}^\top &= \begin{bmatrix} \mathbf{Y}\mathbf{S}_{\text{top}}^\top - \hat{\mathbf{Y}}_{\text{top}} & \mathbf{Y} - \hat{\mathbf{Y}}_{\text{bot}} \end{bmatrix} \begin{bmatrix} \hat{\beta}_{\text{top}}^m \\ \hat{\beta}_{\text{bot}}^m \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{Y}\mathbf{S}_{\text{top}}^\top - \hat{\mathbf{Y}}_{\text{top}} & \mathbf{Y} - \hat{\mathbf{Y}}_{\text{bot}} \end{bmatrix} \begin{bmatrix} \hat{\beta}_{\text{top}}^m \\ \mathbf{I} - \mathbf{S}_{\text{top}}^\top \hat{\beta}_{\text{top}}^m \end{bmatrix} \\
&= (\mathbf{Y}\mathbf{S}_{\text{top}}^\top - \hat{\mathbf{Y}}_{\text{top}} - (\mathbf{Y} - \hat{\mathbf{Y}}_{\text{bot}})\mathbf{S}_{\text{top}}^\top) \hat{\beta}_{\text{top}}^m + \mathbf{Y} - \hat{\mathbf{Y}}_{\text{bot}} \\
&= \mathbf{Y} - \hat{\mathbf{Y}}_{\text{bot}} - (\hat{\mathbf{Y}}_{\text{top}} - \hat{\mathbf{Y}}_{\text{bot}}\mathbf{S}_{\text{top}}^\top) \hat{\beta}_{\text{top}}^m,
\end{aligned} \tag{E.3}$$

which completes the proof. ■

F. Proof of Lemma 2.1

We assume that $\mathbf{X}_{1,\cdot}$ has full rank (i.e., $\text{rank}(\mathbf{X}_{1,\cdot}) = n - m$). In that case the reconciled forecast is

$$\begin{aligned}
\tilde{\mathbf{Y}} &= \hat{\mathbf{Y}}_{\text{bot}} + \mathbf{X}_{1,\cdot} \hat{\beta}_{\text{top}}^m \\
&= \hat{\mathbf{Y}}_{\text{bot}} + \mathbf{X}_{1,\cdot} (\mathbf{X}_{1,\cdot}^\top \mathbf{X}_{1,\cdot})^{-1} \mathbf{X}_{1,\cdot}^\top (\mathbf{Y} - \hat{\mathbf{Y}}_{\text{bot}}) \\
&= \hat{\mathbf{Y}}_{\text{bot}} + \mathbf{H}(\mathbf{Y} - \hat{\mathbf{Y}}_{\text{bot}}),
\end{aligned} \tag{F.1}$$

with $\mathbf{H} = \mathbf{X}_{1,\cdot} (\mathbf{X}_{1,\cdot}^\top \mathbf{X}_{1,\cdot})^{-1} \mathbf{X}_{1,\cdot}^\top$. Both $\mathbf{H}$ and $\mathbf{I} - \mathbf{H}$ are projection matrices, and

$$\begin{aligned}
\tilde{\mathbf{Y}} - \hat{\mathbf{Y}}_{\text{bot}} &= \mathbf{H}(\mathbf{Y} - \hat{\mathbf{Y}}_{\text{bot}}) \\
\mathbf{Y} - \tilde{\mathbf{Y}} &= \mathbf{Y} - \mathbf{H}(\mathbf{Y} - \hat{\mathbf{Y}}_{\text{bot}}) - \hat{\mathbf{Y}}_{\text{bot}} \\
&= (\mathbf{I} - \mathbf{H})(\mathbf{Y} - \hat{\mathbf{Y}}_{\text{bot}}).
\end{aligned} \tag{F.2}$$

For the top levels note that $\mathbf{X}_{1,\cdot} = \hat{\mathbf{Y}}_{\text{top}} - \hat{\mathbf{Y}}_{\text{bot}}\mathbf{S}_{\text{top}}^\top$, and

$$\begin{aligned}
\tilde{\mathbf{Y}}\mathbf{S}_{\text{top}}^\top - \hat{\mathbf{Y}}_{\text{top}} &= \hat{\mathbf{Y}}_{\text{bot}}\mathbf{S}_{\text{top}}^\top + \mathbf{H}(\mathbf{Y} - \hat{\mathbf{Y}}_{\text{bot}})\mathbf{S}_{\text{top}}^\top - \hat{\mathbf{Y}}_{\text{top}} \\
&= \mathbf{H}(\mathbf{Y} - \hat{\mathbf{Y}}_{\text{bot}})\mathbf{S}_{\text{top}}^\top - \mathbf{X}_{1,\cdot}.
\end{aligned} \tag{F.3}$$

Using that $\mathbf{X}_{1,\cdot} = \mathbf{H}\mathbf{X}_{1,\cdot}$, we can write

$$\begin{aligned}
\tilde{\mathbf{Y}}\mathbf{S}_{\text{top}}^\top - \hat{\mathbf{Y}}_{\text{top}} &= \mathbf{H}(\mathbf{Y}\mathbf{S}_{\text{top}}^\top - \hat{\mathbf{Y}}_{\text{bot}}\mathbf{S}_{\text{top}}^\top - \mathbf{X}_{1,\cdot}) \\
&= \mathbf{H}(\mathbf{Y}\mathbf{S}_{\text{top}}^\top - \hat{\mathbf{Y}}_{\text{bot}}\mathbf{S}_{\text{top}}^\top - \hat{\mathbf{Y}}_{\text{top}} + \hat{\mathbf{Y}}_{\text{bot}}\mathbf{S}_{\text{top}}^\top) \\
&= \mathbf{H}(\mathbf{Y}\mathbf{S}_{\text{top}}^\top - \hat{\mathbf{Y}}_{\text{top}}).
\end{aligned} \tag{F.4}$$

As an immediate consequence we can write

$$\begin{aligned} \mathbf{Y} \mathbf{S}_{\text{top}}^\top - \tilde{\mathbf{Y}} \mathbf{S}_{\text{top}}^\top &= \mathbf{Y} \mathbf{S}_{\text{top}}^\top - \mathbf{H}(\mathbf{Y} \mathbf{S}_{\text{top}}^\top - \hat{\mathbf{Y}}_{\text{top}}) - \hat{\mathbf{Y}}_{\text{top}} \\ &= (\mathbf{I} - \mathbf{H})(\mathbf{Y} \mathbf{S}_{\text{top}}^\top - \hat{\mathbf{Y}}_{\text{top}}). \end{aligned} \quad (\text{F.5})$$

Hence,

$$\begin{aligned} \tilde{\mathbf{Y}} \mathbf{S}^\top - \hat{\mathbf{Y}} &= \begin{bmatrix} \mathbf{H}(\mathbf{Y} \mathbf{S}_{\text{top}}^\top - \hat{\mathbf{Y}}_{\text{top}}) & \mathbf{H}(\mathbf{Y} - \hat{\mathbf{Y}}_{\text{bot}}) \end{bmatrix} \\ &= \mathbf{H}(\mathbf{Y} \mathbf{S}^\top - \hat{\mathbf{Y}}) \\ (\mathbf{Y} - \tilde{\mathbf{Y}}) \mathbf{S}^\top &= \begin{bmatrix} \mathbf{Y} \mathbf{S}_{\text{top}}^\top - \tilde{\mathbf{Y}} \mathbf{S}_{\text{top}}^\top & \mathbf{Y} - \tilde{\mathbf{Y}} \end{bmatrix} \\ &= \begin{bmatrix} (\mathbf{I} - \mathbf{H})(\mathbf{Y} \mathbf{S}_{\text{top}}^\top - \hat{\mathbf{Y}}_{\text{top}}) & (\mathbf{I} - \mathbf{H})(\mathbf{Y} - \hat{\mathbf{Y}}_{\text{bot}}) \end{bmatrix} \\ &= (\mathbf{I} - \mathbf{H})(\mathbf{Y} \mathbf{S}^\top - \hat{\mathbf{Y}}), \end{aligned} \quad (\text{F.6})$$

and

$$(\mathbf{Y} \mathbf{S}^\top - \hat{\mathbf{Y}}) = (\mathbf{I} - \mathbf{H})(\mathbf{Y} \mathbf{S}^\top - \hat{\mathbf{Y}}) + \mathbf{H}(\mathbf{Y} \mathbf{S}^\top - \hat{\mathbf{Y}}). \quad (\text{F.7})$$

Orthogonality follows from $\mathbf{H} = \mathbf{H}^\top$ and $\mathbf{H}^2 = \mathbf{H}$. This is true elementwise as we can choose an arbitrary index set I and get

$$\left((\mathbf{Y} \mathbf{S}^\top - \hat{\mathbf{Y}})^\top (\mathbf{Y} \mathbf{S}^\top - \hat{\mathbf{Y}}) \right)_{I,I} = (\mathbf{Y} \mathbf{S}_I^\top - \hat{\mathbf{Y}}_I)^\top (\mathbf{Y} \mathbf{S}_I^\top - \hat{\mathbf{Y}}_I), \quad (\text{F.8})$$

and similarly for the terms on the right-hand side of (F.7). This concludes the proof of Lemma 2.1. ■

G. Proof of Theorem 2.3

Use (F.6) and (B.6) to get

$$\begin{aligned}
\tilde{\mathbf{y}}_I - \hat{\mathbf{y}}_I &= \text{vec}(\mathbf{S}_I \tilde{\mathbf{Y}}^\top - \hat{\mathbf{Y}}_I^\top) \\
&= \text{vec}((\mathbf{S}_I \mathbf{Y}^\top - \hat{\mathbf{Y}}_I^\top) \mathbf{H}) \\
&= (\mathbf{H} \otimes \mathbf{I}_{q_I}) \text{vec}(\mathbf{S}_I \mathbf{Y}^\top - \hat{\mathbf{Y}}_I^\top) \\
&= (\mathbf{H} \otimes \mathbf{I}_{q_I})(\mathbf{y}_I - \hat{\mathbf{y}}_I) \\
\mathbf{y}_I - \tilde{\mathbf{y}}_I &= \text{vec}(\mathbf{S}_I \mathbf{Y}^\top - \mathbf{S}_I \tilde{\mathbf{Y}}_I^\top) \\
&= \text{vec}((\mathbf{S}_I \mathbf{Y}^\top - \hat{\mathbf{Y}}_I^\top)(\mathbf{I} - \mathbf{H})) \\
&= ((\mathbf{I} - \mathbf{H}) \otimes \mathbf{I}_{q_I}) \text{vec}(\mathbf{S}_I \mathbf{Y}^\top - \hat{\mathbf{Y}}_I^\top) \\
&= (\mathbf{I}_{T_{q_I}} - \mathbf{H} \otimes \mathbf{I}_{q_I})(\mathbf{y}_I - \hat{\mathbf{y}}_I).
\end{aligned} \tag{G.1}$$

With $\mathbf{H}_I = \mathbf{H} \otimes \mathbf{I}_{q_I}$ we get

$$\mathbf{y}_I - \hat{\mathbf{y}}_I = \mathbf{H}_I(\mathbf{y}_I - \hat{\mathbf{y}}_I) + (\mathbf{I}_{T_{q_I}} - \mathbf{H}_I)(\mathbf{y}_I - \hat{\mathbf{y}}_I). \tag{G.2}$$

Since $\mathbf{H}$ is a projection matrix we get $\mathbf{H}_I^2 = (\mathbf{H} \otimes \mathbf{I}_{q_I})(\mathbf{H} \otimes \mathbf{I}_{q_I}) = \mathbf{H}^2 \otimes \mathbf{I}_{q_I}^2 = \mathbf{H} \otimes \mathbf{I}_{q_I}$ and $\mathbf{H}_I^\top = (\mathbf{H} \otimes \mathbf{I}_{q_I})^\top = \mathbf{H}^\top \otimes \mathbf{I}_{q_I}^\top = \mathbf{H} \otimes \mathbf{I}_{q_I} = \mathbf{H}_I$. Orthogonality follows from $(\mathbf{I}_{T_{q_I}} - \mathbf{H}_I)\mathbf{H}_I = \mathbf{0}$. Thus, we have

$$\begin{aligned}
\|\mathbf{y}_I - \hat{\mathbf{y}}_I\|^2 &= (\mathbf{y}_I - \hat{\mathbf{y}}_I)^\top \mathbf{H}_I(\mathbf{y}_I - \hat{\mathbf{y}}_I) + (\mathbf{y}_I - \hat{\mathbf{y}}_I)^\top (\mathbf{I}_{T_{q_I}} - \mathbf{H}_I)(\mathbf{y}_I - \hat{\mathbf{y}}_I) \\
&= \|\mathbf{y}_I - \tilde{\mathbf{y}}_I\|^2 + \|\tilde{\mathbf{y}}_I - \hat{\mathbf{y}}_I\|^2,
\end{aligned} \tag{G.3}$$

which is (30), and (31) follows directly. This concludes the proof of Theorem 2.3. $\blacksquare$

H. Proof of Corollary 2.2

We consider the model

$$\mathbf{y}_t - \hat{\mathbf{y}}_{\text{bot},t} = \left[\mathbf{I}_m \otimes (\hat{\mathbf{y}}_{\text{top},t}^\top - \hat{\mathbf{y}}_{\text{bot},t}^\top \mathbf{S}_{\text{top}}^\top) \right] \boldsymbol{\beta}_{\text{top}} + \boldsymbol{\epsilon}_t; \quad \boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma}_r). \tag{H.1}$$

Multiplying with a fixed vector $\mathbf{v}$ and setting $\mathbf{X}_{1,t} = \hat{\mathbf{y}}_{\text{top},t}^\top - \hat{\mathbf{y}}_{\text{bot},t}^\top \mathbf{S}_{\text{top}}^\top$, we get

$$\mathbf{v}^\top (\mathbf{y}_t - \hat{\mathbf{y}}_{\text{bot},t}) = [v_1 \mathbf{X}_{1,t} \quad v_2 \mathbf{X}_{1,t} \quad \dots \quad v_m \mathbf{X}_{1,t}] \boldsymbol{\beta}_{\text{top}} + \boldsymbol{\epsilon}_t; \quad \mathbf{v}^\top \boldsymbol{\epsilon}_t \sim N(0, \boldsymbol{\Sigma}_r). \tag{H.2}$$

Setting $\gamma = \sum_{i=1}^m v_i \beta_{\text{top},i}$, $z_t = \mathbf{v}^\top (\mathbf{y}_t - \hat{\mathbf{y}}_{\text{bot},t})$, $u_t = \mathbf{v}^\top \epsilon_t$, and $\sigma_v^2 = \mathbf{v}^\top \Sigma_r \mathbf{v}$, we get

$$z_t = \mathbf{X}_{1,t} \gamma + u_t; \quad u_t \sim N(0, \sigma_v^2). \quad (\text{H.3})$$

Written in matrix-vector notation this is

$$\mathbf{z} = \mathbf{X}_{1,\cdot} \gamma + \mathbf{u}; \quad \mathbf{u} \sim N(0, \sigma_v^2 \mathbf{I}), \quad (\text{H.4})$$

which is the usual general linear model for which variance decomposition and Cochran's Theorem imply the χ^2 -distribution (see e.g. Madsen and Thyregod, 2011). For completeness we write out the chain of projection. The model can be written as

$$\begin{aligned} \mathbf{u} &= \mathbf{z} - \mathbf{X}_{1,\cdot} \gamma \\ &= \mathbf{z} - \mathbf{X}_{1,\cdot} \hat{\gamma} + \mathbf{X}_{1,\cdot} \hat{\gamma} - \mathbf{X}_{1,\cdot} \gamma \\ &= (\mathbf{I} - \mathbf{H}) \mathbf{z} + \mathbf{H} \mathbf{z} - \mathbf{X}_{1,\cdot} \gamma. \end{aligned} \quad (\text{H.5})$$

By definition $\frac{\mathbf{u}^\top \mathbf{u}}{\sigma_v^2} \sim \chi^2(T)$. For the second term, centrality of the estimator $\hat{\gamma}$ implies that $E[(\mathbf{I} - \mathbf{H}) \mathbf{z}] = \mathbf{0}$. Since $(\mathbf{I} - \mathbf{H})$ is a projection matrix there exists an orthogonal basis that spans the space of $(\mathbf{I} - \mathbf{H}) \mathbf{z}$ and, hence, by Lemma B.1, $\frac{\mathbf{z}^\top (\mathbf{I} - \mathbf{H}) \mathbf{z}}{\sigma_v^2} \sim \chi^2(T - n + m)$. Additionally, if $\gamma = \mathbf{0}$, then $\frac{\mathbf{z}^\top \mathbf{H} \mathbf{z}}{\sigma_v^2} \sim \chi^2(n - m)$. Since $\mathbf{v}$ is arbitrary, Lemma B.1 implies that

$$(\mathbf{Y} - \tilde{\mathbf{Y}})^\top (\mathbf{Y} - \tilde{\mathbf{Y}}) \sim W_m(\Sigma_r, T - (n - m)). \quad (\text{H.6})$$

This holds under the model assumptions (i.e., also when $\beta_{\text{top}} \neq \mathbf{0}$). If, in addition, $\gamma = \mathbf{0}$ for all $\mathbf{v}$, implying that $\beta_{\text{top}} = \mathbf{0}$, then

$$(\tilde{\mathbf{Y}} - \hat{\mathbf{Y}})^\top (\tilde{\mathbf{Y}} - \hat{\mathbf{Y}}) \sim W_m(\Sigma_r, n - m). \quad (\text{H.7})$$

which concludes the proof of Corollary 2.2. ■

I. Proof of Theorem 3.1

The usual shrinkage estimator for the variance–covariance matrix is

$$\hat{\Sigma}_s = (1 - \lambda) \hat{\Sigma}_h + \lambda \hat{\Sigma}_h^d, \quad (\text{I.1})$$

with $\hat{\Sigma}_h = \frac{1}{T} \left(\mathbf{Y} \mathbf{S}^\top - \hat{\mathbf{Y}} \right)^\top \left(\mathbf{Y} \mathbf{S}^\top - \hat{\mathbf{Y}} \right)$. Using Theorem 1 of Wickramasuriya et al. (2019) and (D.11)–(D.12) we can write

$$\mathbf{J} \hat{\Sigma}_s \mathbf{U} = -\frac{1-\lambda}{T} \left(\mathbf{Y} - \hat{\mathbf{Y}}_{\text{bot}} \right)^\top \left(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{\text{bot}} \mathbf{S}_{\text{top}}^\top \right) + \lambda \mathbf{J} \hat{\Sigma}_h^{\text{d}} \mathbf{U}, \quad (\text{I.2})$$

where

$$\begin{aligned} \mathbf{J} \hat{\Sigma}_h^{\text{d}} \mathbf{U} &= [\mathbf{0}_{m \times n-m} \quad \mathbf{I}_m] \begin{bmatrix} \hat{\Sigma}_{h,\text{top}}^{\text{d}} & \mathbf{0} \\ \mathbf{0} & \hat{\Sigma}_{h,\text{bot}}^{\text{d}} \end{bmatrix} \begin{bmatrix} \mathbf{I}_{n-m} \\ -\mathbf{S}_{\text{top}}^\top \end{bmatrix} \\ &= -\hat{\Sigma}_{h,\text{bot}}^{\text{d}} \mathbf{S}_{\text{top}}^\top. \end{aligned} \quad (\text{I.3})$$

Also, using (D.11), we can write

$$\mathbf{U}^\top \hat{\Sigma}_s \mathbf{U} = \frac{1-\lambda}{T} \left(\mathbf{Y} - \hat{\mathbf{Y}}_{\text{bot}} \right)^\top \left(\mathbf{Y} - \hat{\mathbf{Y}}_{\text{bot}} \right) + \lambda \mathbf{U}^\top \hat{\Sigma}_h^{\text{d}} \mathbf{U}, \quad (\text{I.4})$$

and

$$\begin{aligned} \mathbf{U}^\top \hat{\Sigma}_h^{\text{d}} \mathbf{U} &= \begin{bmatrix} \mathbf{I}_{n-m} & -\mathbf{S}_{\text{top}} \end{bmatrix} \begin{bmatrix} \hat{\Sigma}_{h,\text{top}}^{\text{d}} & \mathbf{0} \\ \mathbf{0} & \hat{\Sigma}_{h,\text{bot}}^{\text{d}} \end{bmatrix} \begin{bmatrix} \mathbf{I}_{n-m} \\ -\mathbf{S}_{\text{top}}^\top \end{bmatrix} \\ &= \hat{\Sigma}_{h,\text{top}}^{\text{d}} + \mathbf{S}_{\text{top}} \hat{\Sigma}_{h,\text{bot}}^{\text{d}} \mathbf{S}_{\text{top}}^\top. \end{aligned} \quad (\text{I.5})$$

Hence,

$$\begin{aligned} \mathbf{P}(\lambda) &= \mathbf{J} - \mathbf{J} \hat{\Sigma}_s \mathbf{U} \left(\mathbf{U}^\top \hat{\Sigma}_s \mathbf{U} \right)^{-1} \mathbf{U}^\top \\ &= \mathbf{J} + \left(\frac{1-\lambda}{T} \left(\mathbf{Y} - \hat{\mathbf{Y}}_{\text{bot}} \right)^\top \left(\hat{\mathbf{Y}}_{\text{top}} - \hat{\mathbf{Y}}_{\text{bot}} \mathbf{S}_{\text{top}}^\top \right) + \lambda \hat{\Sigma}_{h,\text{bot}}^{\text{d}} \mathbf{S}_{\text{top}}^\top \right) \times \\ &\quad \left(\frac{1-\lambda}{T} \left(\hat{\mathbf{Y}}_{\text{top}} - \hat{\mathbf{Y}}_{\text{bot}} \mathbf{S}_{\text{top}}^\top \right)^\top \left(\hat{\mathbf{Y}}_{\text{top}} - \hat{\mathbf{Y}}_{\text{bot}} \mathbf{S}_{\text{top}}^\top \right) + \lambda \left(\hat{\Sigma}_{h,\text{top}}^{\text{d}} + \mathbf{S}_{\text{top}} \hat{\Sigma}_{h,\text{bot}}^{\text{d}} \mathbf{S}_{\text{top}}^\top \right) \right)^{-1} \mathbf{U}^\top. \end{aligned} \quad (\text{I.6})$$

Now consider the model

$$\mathbf{y}_t - \hat{\mathbf{y}}_{\text{bot},t} = \left[\mathbf{I}_m \otimes \hat{\mathbf{g}}_{\text{top},t}^\top - \mathbf{I}_m \otimes \left(\hat{\mathbf{g}}_{\text{bot},t}^\top \mathbf{S}_{\text{top}}^\top \right) \right] \boldsymbol{\beta}_{\text{top}} + \boldsymbol{\epsilon}_t, \quad (\text{I.7})$$

or

$$\mathbf{y} - \hat{\mathbf{y}}_{\text{bot}} = \mathbf{X} \boldsymbol{\beta}_{\text{top}} + \boldsymbol{\epsilon}, \quad (\text{I.8})$$

with $\epsilon \sim N(\mathbf{0}, \mathbf{I}_T \otimes \Sigma_r)$ and $\beta_{\text{top}} \sim N_{m(n-m)}(\beta_{0,\text{top}}, \Sigma_\beta)$. Its log-posterior density is

$$l_{\text{MAP}}(\beta_{\text{top}}, \Sigma_r) \propto -\frac{1}{2} \log |\Sigma| - \frac{1}{2} \left(\mathbf{y} - \hat{\mathbf{Y}}_{\text{bot}} - \mathbf{X} \beta_{\text{top}} \right)^\top \Sigma^{-1} \left(\mathbf{y} - \hat{\mathbf{Y}}_{\text{bot}} - \mathbf{X} \beta_{\text{top}} \right) - \frac{1}{2} (\beta_{\text{top}} - \beta_{0T})^\top \Sigma_\beta^{-1} (\beta_{\text{top}} - \beta_{0,\text{top}}) - \frac{1}{2} \log |\Sigma_\beta|, \quad (\text{I.9})$$

and, consequently, the MAP estimate of β_{top} is

$$\hat{\beta}_{\text{top}} = \left(\mathbf{X}^\top (\mathbf{I}_T \otimes \Sigma_r)^{-1} \mathbf{X} + \Sigma_\beta^{-1} \right)^{-1} \left(\mathbf{X}^\top (\mathbf{I}_T \otimes \Sigma_r)^{-1} (\mathbf{y} - \hat{\mathbf{Y}}_{\text{bot}}) + \Sigma_\beta^{-1} \beta_{0,\text{top}} \right). \quad (\text{I.10})$$

Choosing $\Sigma_\beta = \Sigma_r \otimes \Sigma_{0,\beta}$, we go through each term. Since $\mathbf{x}_{t,i} = \mathbf{x}_{t,j}$, using (C.7), we get

$$\left(\mathbf{X}^\top (\mathbf{I}_T \otimes \Sigma_r)^{-1} \mathbf{X} + \Sigma_\beta^{-1} \right)^{-1} = \Sigma_r \otimes \left(\mathbf{X}_{1,\cdot}^\top \mathbf{X}_{1,\cdot} + \Sigma_{0,\beta}^{-1} \right)^{-1}, \quad (\text{I.11})$$

and using, (C.10)

$$\begin{aligned} \left(\mathbf{X}^\top (\mathbf{I}_T \otimes \Sigma_r)^{-1} (\mathbf{y} - \hat{\mathbf{Y}}_{\text{bot}}) \right) &= \left(\mathbf{X}^\top \text{vec}(\Sigma_r^{-1} (\mathbf{Y} - \hat{\mathbf{Y}}_{\text{bot}})^\top) \right) \\ &= \left(\mathbf{I}_m \otimes \mathbf{X}_{1,\cdot}^\top \right) \text{vec} \left((\mathbf{Y} - \hat{\mathbf{Y}}_{\text{bot}}) \Sigma_r^{-1} \right) \\ &= \text{vec} \left(\mathbf{X}_{1,\cdot}^\top (\mathbf{Y} - \hat{\mathbf{Y}}_{\text{bot}}) \Sigma_r^{-1} \right). \end{aligned} \quad (\text{I.12})$$

It follows from the definition of the prior and (B.3) that

$$\Sigma_\beta^{-1} \beta_{0,\text{top}} = (\Sigma_r^{-1} \otimes \Sigma_{0,\beta}^{-1}) \beta_{0,\text{top}}. \quad (\text{I.13})$$

Hence (I.10), can be written as

$$\hat{\beta}_{\text{top}} = \left(\Sigma_r \otimes \left(\mathbf{X}_{1,\cdot}^\top \mathbf{X}_{1,\cdot} + \Sigma_{0,\beta}^{-1} \right)^{-1} \right) \left(\text{vec} \left(\mathbf{X}_{1,\cdot}^\top (\mathbf{Y} - \hat{\mathbf{Y}}_{\text{bot}}) \Sigma_r^{-1} \right) + (\Sigma_r^{-1} \otimes \Sigma_{0,\beta}^{-1}) \beta_{0,\text{top}} \right). \quad (\text{I.14})$$

Multiplying terms (using $\mathbf{C} = \left(\mathbf{X}_{1,\cdot}^\top \mathbf{X}_{1,\cdot} + \Sigma_{0,\beta}^{-1} \right)^{-1}$) we get

$$(\Sigma_r \otimes \mathbf{C}) \text{vec} \left(\mathbf{X}_{1,\cdot}^\top (\mathbf{Y} - \hat{\mathbf{Y}}_{\text{bot}}) \Sigma_r^{-1} \right) = \text{vec} \left(\mathbf{C} \mathbf{X}_{1,\cdot}^\top (\mathbf{Y} - \hat{\mathbf{Y}}_{\text{bot}}) \right) \quad (\text{I.15})$$

and

$$\begin{aligned} (\Sigma_r \otimes \mathbf{C}) (\Sigma_r^{-1} \otimes \Sigma_{0,\beta}^{-1}) \beta_{0,\text{top}} &= \left(\mathbf{I}_m \otimes \left(\mathbf{C} \Sigma_{0,\beta}^{-1} \right) \right) \beta_{0,\text{top}} \\ &= \text{vec} \left(\mathbf{C} \Sigma_{0,\beta}^{-1} \beta_{0,\text{top}}^m \right). \end{aligned} \quad (\text{I.16})$$

Inserting (I.15)–(I.16) in (I.14) we get

$$\hat{\beta}_{\text{top}} = \text{vec} \left(\left(\mathbf{X}_{1,\cdot}^\top \mathbf{X}_{1,\cdot} + \Sigma_{0,\beta}^{-1} \right)^{-1} \left(\mathbf{X}_{1,\cdot}^\top (\mathbf{Y} - \hat{\mathbf{Y}}_{\text{bot}}) + \Sigma_{0,\beta}^{-1} \beta_{0,\text{top}}^m \right) \right). \quad (\text{I.17})$$

With $\Sigma_{0,\beta} = \frac{1-\lambda}{\lambda T} (\hat{\Sigma}_{h,\text{top}}^d + \mathbf{S}_{\text{top}} \hat{\Sigma}_{h,\text{bot}}^d \mathbf{S}_{\text{top}}^\top)^{-1}$ and $\mathbf{X}_{1,\cdot} = \hat{\mathbf{Y}}_{\text{top}} - \hat{\mathbf{Y}}_{\text{bot}} \mathbf{S}_{\text{top}}^\top$, we get

$$\begin{aligned} \hat{\beta}_{\text{top}} &= \text{vec} \left[\left((\hat{\mathbf{Y}}_{\text{top}} - \hat{\mathbf{Y}}_{\text{bot}} \mathbf{S}_{\text{top}}^\top)^\top (\hat{\mathbf{Y}}_{\text{top}} - \hat{\mathbf{Y}}_{\text{bot}} \mathbf{S}_{\text{top}}^\top) + \frac{\lambda T}{1-\lambda} (\hat{\Sigma}_{h,\text{top}}^d + \mathbf{S}_{\text{top}} \hat{\Sigma}_{h,\text{bot}}^d \mathbf{S}_{\text{top}}^\top) \right)^{-1} \times \right. \\ &\quad \left. \left((\hat{\mathbf{Y}}_{\text{top}} - \hat{\mathbf{Y}}_{\text{bot}} \mathbf{S}_{\text{top}}^\top)^\top (\mathbf{Y} - \hat{\mathbf{Y}}_{\text{bot}}) + \frac{\lambda T}{1-\lambda} (\hat{\Sigma}_{h,\text{top}}^d + \mathbf{S}_{\text{top}} \hat{\Sigma}_{h,\text{bot}}^d \mathbf{S}_{\text{top}}^\top) (\beta_{0,\text{top}}^m)^\top \right) \right] \\ &= \text{vec} \left[\left((1-\lambda) (\hat{\mathbf{Y}}_{\text{top}} - \hat{\mathbf{Y}}_{\text{bot}} \mathbf{S}_{\text{top}}^\top)^\top (\hat{\mathbf{Y}}_{\text{top}} - \hat{\mathbf{Y}}_{\text{bot}} \mathbf{S}_{\text{top}}^\top) + \lambda T (\hat{\Sigma}_{h,\text{top}}^d + \mathbf{S}_{\text{top}} \hat{\Sigma}_{h,\text{bot}}^d \mathbf{S}_{\text{top}}^\top) \right)^{-1} \times \right. \\ &\quad \left. \left((1-\lambda) (\hat{\mathbf{Y}}_{\text{top}} - \hat{\mathbf{Y}}_{\text{bot}} \mathbf{S}_{\text{top}}^\top)^\top (\mathbf{Y} - \hat{\mathbf{Y}}_{\text{bot}}) + \lambda T (\hat{\Sigma}_{h,\text{top}}^d + \mathbf{S}_{\text{top}} \hat{\Sigma}_{h,\text{bot}}^d \mathbf{S}_{\text{top}}^\top) (\beta_{0,\text{top}}^m)^\top \right) \right]. \end{aligned} \quad (\text{I.18})$$

Hence, if we choose $\beta_{0,\text{top}}^m = \hat{\Sigma}_{h,\text{bot}}^d \mathbf{S}_{\text{top}}^\top (\hat{\Sigma}_{h,\text{top}}^d + \mathbf{S}_{\text{top}} \hat{\Sigma}_{h,\text{bot}}^d \mathbf{S}_{\text{top}}^\top)^{-1}$, we get

$$\begin{aligned} \hat{\beta}_{\text{top}}^m &= \left((1-\lambda) (\hat{\mathbf{Y}}_{\text{top}} - \hat{\mathbf{Y}}_{\text{bot}} \mathbf{S}_{\text{top}}^\top)^\top (\hat{\mathbf{Y}}_{\text{top}} - \hat{\mathbf{Y}}_{\text{bot}} \mathbf{S}_{\text{top}}^\top) + \lambda T (\hat{\Sigma}_{h,\text{top}}^d + \mathbf{S}_{\text{top}} \hat{\Sigma}_{h,\text{bot}}^d \mathbf{S}_{\text{top}}^\top) \right)^{-1} \times \\ &\quad \left((1-\lambda) (\hat{\mathbf{Y}}_{\text{top}} - \hat{\mathbf{Y}}_{\text{bot}} \mathbf{S}_{\text{top}}^\top)^\top (\mathbf{Y} - \hat{\mathbf{Y}}_{\text{bot}}) + \lambda T \mathbf{S}_{\text{top}} \hat{\Sigma}_{h,\text{bot}}^d \right). \end{aligned} \quad (\text{I.19})$$

Finally, using the linear constraints $\beta_{\text{bot}}^m = \mathbf{I}_m - \mathbf{S}_{\text{top}}^\top \beta_{\text{top}}^m$ and (I.6), we get

$$\begin{aligned} \begin{bmatrix} \hat{\beta}_{\text{top}}^m \\ \hat{\beta}_{\text{bot}}^m \end{bmatrix} &= \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} + \begin{bmatrix} \mathbf{I} \\ -\mathbf{S}_{\text{top}}^\top \end{bmatrix} \hat{\beta}_{\text{top}}^m \\ &= \mathbf{P}^\top(\lambda). \end{aligned} \quad (\text{I.20})$$

This is (41) and concludes the proof of Theorem 3.1. ■

J. Proof of Corollary 3.1

The log-posterior density wrt. Σ_r is

$$\begin{aligned} l_{\text{MAP}}(\Sigma_r) &\propto -\frac{T}{2} \log |\Sigma_r| - \frac{1}{2} \left(\mathbf{y} - \hat{\mathbf{Y}}_{\text{bot}} - \mathbf{X} \beta_{\text{top}} \right)^\top (\mathbf{I}_T \otimes \Sigma_r^{-1}) \left(\mathbf{y} - \hat{\mathbf{Y}}_{\text{bot}} - \mathbf{X} \beta_{\text{top}} \right) - \\ &\quad \frac{1}{2} (\beta_{\text{top}} - \beta_{0,\text{top}})^\top (\Sigma_r^{-1} \otimes \Sigma_{0,\beta}^{-1}) (\beta_{\text{top}} - \beta_{0,\text{top}}) - \frac{n-m}{2} \log |\Sigma_r|. \end{aligned} \quad (\text{J.1})$$

The first two terms and the last term are treated in Appendix C. The only term in (J.1) that we have not treated in the previous is the derivative of the third term. Using (B.16) we get

$$\begin{aligned} (\beta_{\text{top}} - \beta_{0,\text{top}})^\top (\Sigma_r^{-1} \otimes \Sigma_{0,\beta}^{-1}) (\beta_{\text{top}} - \beta_{0,\text{top}}) &= \text{Tr} \left((\beta_{\text{top}}^m - \beta_{0,\text{top}}^m)^\top \Sigma_{0,\beta}^{-1} (\beta_{\text{top}}^m - \beta_{0,\text{top}}^m) \Sigma_r^{-1} \right) \\ &= \text{Tr} \left((\beta_{\text{top}}^m - \beta_{0,\text{top}}^m) \Sigma_r^{-1} (\beta_{\text{top}}^m - \beta_{0,\text{top}}^m)^\top \Sigma_{0,\beta}^{-1} \right), \end{aligned} \quad (\text{J.2})$$

and using (B.14) we get

$$\frac{\partial (\beta_{\text{top}} - \beta_{0,\text{top}})^\top (\Sigma_r^{-1} \otimes \Sigma_{0,\beta}^{-1}) (\beta_{\text{top}} - \beta_{0,\text{top}})}{\partial \Sigma_r} = -\Sigma_r^{-1} (\beta_{\text{top}}^m - \beta_{0,\text{top}}^m)^\top \Sigma_{0,\beta}^{-1} (\beta_{\text{top}}^m - \beta_{0,\text{top}}^m) \Sigma_r^{-1}. \quad (\text{J.3})$$

The derivative of l_{MAP} wrt. Σ_r is therefore

$$\begin{aligned} \frac{\partial l_{\text{MAP}}(\Sigma_r)}{\partial \Sigma_r} &= -\frac{T}{2} \Sigma_r^{-1} + \frac{1}{2} \Sigma_r^{-1} \sum_{t=1}^T \mathbf{e}_t \mathbf{e}_t^\top \Sigma_r^{-1} + \\ &\quad \frac{1}{2} \Sigma_r^{-1} (\hat{\beta}_{\text{top}}^m - \beta_{0,\text{top}}^m)^\top \Sigma_{0,\beta}^{-1} (\hat{\beta}_{\text{top}}^m - \beta_{0,\text{top}}^m) \Sigma_r^{-1} - \frac{n-m}{2} \Sigma_r^{-1} \end{aligned} \quad (\text{J.4})$$

and, hence, the MAP estimate of Σ_r is

$$\begin{aligned} \hat{\Sigma}_r &= \frac{1}{T+n-m} \left(\sum_{t=1}^T \mathbf{e}_t \mathbf{e}_t^\top + (\hat{\beta}_{\text{top}}^m - \beta_{0,\text{top}}^m)^\top \Sigma_{0,\beta}^{-1} (\hat{\beta}_{\text{top}}^m - \beta_{0,\text{top}}^m) \right) \\ &= \frac{1}{T+n-m} \left(\sum_{t=1}^T \mathbf{e}_t \mathbf{e}_t^\top + \right. \\ &\quad \left. \frac{\lambda T}{1-\lambda} (\hat{\beta}_{\text{top}}^m - \beta_{0,\text{top}}^m)^\top (\hat{\Sigma}_{h,\text{top}}^d + \mathbf{S}_{\text{top}} \hat{\Sigma}_{h,\text{bot}}^d \mathbf{S}_{\text{top}}^\top) (\hat{\beta}_{\text{top}}^m - \beta_{0,\text{top}}^m) \right) \\ &= \frac{T/(1-\lambda)}{T+n-m} \left(\frac{1-\lambda}{T} \sum_{t=1}^T \mathbf{e}_t \mathbf{e}_t^\top + \right. \\ &\quad \left. \lambda (\hat{\beta}_{\text{top}}^m - \beta_{0,\text{top}}^m)^\top (\hat{\Sigma}_{h,\text{top}}^d + \mathbf{S}_{\text{top}} \hat{\Sigma}_{h,\text{bot}}^d \mathbf{S}_{\text{top}}^\top) (\hat{\beta}_{\text{top}}^m - \beta_{0,\text{top}}^m) \right). \end{aligned} \quad (\text{J.5})$$

Using (the proof of) Corollary 2.1 (Appendix E) and inserting $\beta_{0,\text{top}}$ we can rewrite this as

$$\begin{aligned} \hat{\Sigma}_r &= \frac{T/(1-\lambda)}{T+n-m} \left((1-\lambda) \mathbf{P}(\lambda) \hat{\Sigma}_h \mathbf{P}^\top(\lambda) + \lambda \left((\hat{\beta}_{\text{top}}^m)^\top (\hat{\Sigma}_{h,\text{top}}^d + \mathbf{S}_{\text{top}} \hat{\Sigma}_{h,\text{bot}}^d \mathbf{S}_{\text{top}}^\top) \hat{\beta}_{\text{top}}^m - \right. \right. \\ &\quad \left. \left. (\hat{\beta}_{\text{top}}^m)^\top \mathbf{S}_{\text{top}} \hat{\Sigma}_{h,\text{bot}}^d - \hat{\Sigma}_{h,\text{bot}}^d \mathbf{S}_{\text{top}}^\top \hat{\beta}_{\text{top}}^m + \hat{\Sigma}_{h,\text{bot}}^d \mathbf{S}_{\text{top}}^\top (\hat{\Sigma}_{h,\text{top}}^d + \mathbf{S}_{\text{top}} \hat{\Sigma}_{h,\text{bot}}^d \mathbf{S}_{\text{top}}^\top)^{-1} \mathbf{S}_{\text{top}} \hat{\Sigma}_{h,\text{bot}}^d \right) \right). \end{aligned} \quad (\text{J.6})$$

With $\hat{\Sigma}_s = (1-\lambda) \hat{\Sigma}_h + \lambda \hat{\Sigma}_h^d$, we have $(1-\lambda) \mathbf{P}(\lambda) \hat{\Sigma}_h \mathbf{P}^\top(\lambda) = \mathbf{P}(\lambda) \hat{\Sigma}_s \mathbf{P}^\top(\lambda) - \lambda \mathbf{P}(\lambda) \hat{\Sigma}_h^d \mathbf{P}^\top(\lambda)$,

where

$$\begin{aligned}
P(\lambda) \Sigma_h^d P^\top(\lambda) &= \begin{bmatrix} \hat{\beta}_{\text{top}}^m \\ I - S_{\text{top}}^\top \hat{\beta}_{\text{top}}^m \end{bmatrix}^\top \begin{bmatrix} \Sigma_{h,T}^d & \mathbf{0} \\ \mathbf{0} & \hat{\Sigma}_{h,\text{bot}}^d \end{bmatrix} \begin{bmatrix} \hat{\beta}_{\text{top}}^m \\ I - S_{\text{top}}^\top \hat{\beta}_{\text{top}}^m \end{bmatrix} \\
&= (\hat{\beta}_{\text{top}}^m)^\top \Sigma_{h,T}^d \hat{\beta}_{\text{top}}^m + \left(I - (\hat{\beta}_{\text{top}}^m)^\top S_{\text{top}} \right) \hat{\Sigma}_{h,\text{bot}}^d \left(I - S_{\text{top}}^\top \hat{\beta}_{\text{top}}^m \right) \\
&= (\hat{\beta}_{\text{top}}^m)^\top \Sigma_{h,T}^d \hat{\beta}_{\text{top}}^m + \hat{\Sigma}_{h,\text{bot}}^d - \hat{\Sigma}_{h,\text{bot}}^d S_{\text{top}}^\top \hat{\beta}_{\text{top}}^m - \\
&\quad (\hat{\beta}_{\text{top}}^m)^\top S_{\text{top}} \hat{\Sigma}_{h,\text{bot}}^d + (\hat{\beta}_{\text{top}}^m)^\top S_{\text{top}} \hat{\Sigma}_{h,\text{bot}}^d S_{\text{top}}^\top \hat{\beta}_{\text{top}}^m \\
&= (\hat{\beta}_{\text{top}}^m)^\top \left(\Sigma_{h,T}^d + S_{\text{top}} \hat{\Sigma}_{h,\text{bot}}^d S_{\text{top}}^\top \right) \hat{\beta}_{\text{top}}^m + \hat{\Sigma}_{h,\text{bot}}^d - \\
&\quad \hat{\Sigma}_{h,\text{bot}}^d S_{\text{top}}^\top (\hat{\beta}_{\text{top}}^m)^\top - (\hat{\beta}_{\text{top}}^m)^\top S_{\text{top}} \hat{\Sigma}_{h,\text{bot}}^d.
\end{aligned} \tag{J.7}$$

It follows that the MAP estimate of Σ_r can be written as (using Petersen and Pedersen, 2006, eq. (156) for the second equality)

$$\begin{aligned}
\hat{\Sigma}_r &= \frac{T/(1-\lambda)}{T+n-m} \left(P(\lambda) \hat{\Sigma}_s P^\top(\lambda) - \lambda \hat{\Sigma}_{h,B}^d + \right. \\
&\quad \left. \lambda \hat{\Sigma}_{h,\text{bot}}^d S_{\text{top}}^\top (\hat{\Sigma}_{h,\text{top}}^d + S_{\text{top}} \hat{\Sigma}_{h,\text{bot}}^d S_{\text{top}}^\top)^{-1} S_{\text{top}} \hat{\Sigma}_{h,\text{bot}}^d \right) \\
&= \frac{T/(1-\lambda)}{T+n-m} \left(P(\lambda) \hat{\Sigma}_s P^\top(\lambda) - \lambda \left((\hat{\Sigma}_{h,\text{bot}}^d)^{-1} + S_{\text{top}}^\top (\hat{\Sigma}_{h,\text{top}}^d)^{-1} S_{\text{top}} \right)^{-1} \right).
\end{aligned} \tag{J.8}$$

This is (49) and concludes the proof of Corollary 3.1. ■

K. Proof of Theorem 3.2

If we assume that $\Sigma_r \sim \mathcal{W}^{-1}(\Psi, v)$ and $v > m - 1$, then the log-posterior density becomes

$$\begin{aligned}
l_{\text{MAP}}(\Sigma_r) &\propto -\frac{1}{2} \left(\mathbf{y} - \hat{\mathbf{Y}}_{\text{bot}} - \mathbf{X} \hat{\beta}_{\text{top}} \right)^\top (I_T \otimes \Sigma_r^{-1}) \left(\mathbf{y} - \hat{\mathbf{Y}}_{\text{bot}} - \mathbf{X} \hat{\beta}_{\text{top}} \right) - \\
&\quad \frac{1}{2} (\beta_{\text{top}} - \beta_{0,\text{top}})^\top \left(\Sigma_r^{-1} \otimes \Sigma_{0,\beta}^{-1} \right) (\beta_{\text{top}} - \beta_{0,\text{top}}) - \\
&\quad \frac{n-m}{2} \log |\Sigma_r| - \frac{T}{2} \log |\Sigma_r| - \frac{v+m+1}{2} |\Sigma_r| - \frac{1}{2} \text{Tr}(\Psi \Sigma_r^{-1}),
\end{aligned} \tag{K.1}$$

and the derivative wrt. Σ_r is

$$\begin{aligned}
\frac{\partial l_{\text{MAP}}(\Sigma_r)}{\partial \Sigma_r} &= -\frac{T+n+v+1}{2} \Sigma_r^{-1} + \frac{1}{2} \Sigma_r^{-1} \sum_{t=1}^T \mathbf{e}_t \mathbf{e}_t^\top \Sigma_r^{-1} + \\
&\quad \frac{1}{2} \Sigma_r^{-1} \left(\hat{\beta}_{\text{top}}^m - \beta_{0,\text{top}}^m \right) \Sigma_{0,\beta}^{-1} \left(\hat{\beta}_{\text{top}}^m - \beta_{0,\text{top}}^m \right)^\top \Sigma_r^{-1} - \frac{1}{2} \Sigma_r^{-1} \Psi \Sigma_r^{-1}.
\end{aligned} \tag{K.2}$$

Consequently, using (J.8) the MAP estimate of Σ_r is

$$\begin{aligned}\hat{\Sigma}_{r,\text{MAP}} &= \frac{\left(\sum_{t=1}^T \mathbf{e}_t \mathbf{e}_t^\top + \left(\hat{\beta}_{\text{top}}^m - \beta_{0,\text{top}}^m\right) \Sigma_{0,\beta}^{-1} \left(\hat{\beta}_{\text{top}}^m - \beta_{0,\text{top}}^m\right)^\top + \Psi\right)}{T + n + v + 1} \\ &= \frac{T/(1-\lambda)}{T + n + v + 1} \left(\mathbf{P}(\lambda) \hat{\Sigma}_s \mathbf{P}^\top(\lambda) - \right. \\ &\quad \left. \lambda \left(\left(\hat{\Sigma}_{h,\text{bot}}^{\text{d}} \right)^{-1} + \mathbf{S}_{\text{top}}^\top \left(\hat{\Sigma}_{h,\text{top}}^{\text{d}} \right)^{-1} \mathbf{S}_{\text{top}} \right)^{-1} + \frac{1-\lambda}{T} \Psi \right).\end{aligned}\tag{K.3}$$

If we set $\Psi = \frac{T\lambda}{1-\lambda} \left(\left(\hat{\Sigma}_{h,\text{bot}}^{\text{d}} \right)^{-1} + \mathbf{S}_{\text{top}}^\top \left(\hat{\Sigma}_{h,\text{top}}^{\text{d}} \right)^{-1} \mathbf{S}_{\text{top}} \right)^{-1}$, then the estimate is

$$\hat{\Sigma}_{r,\text{MAP}} = \frac{T/(1-\lambda)}{T + n + v + 1} \mathbf{P}(\lambda) \hat{\Sigma}_s \mathbf{P}^\top(\lambda).\tag{K.4}$$

If, in addition, we set $v = \frac{\lambda}{1-\lambda}T - (n+1)$, then the estimate becomes

$$\hat{\Sigma}_{r,\text{MAP}} = \mathbf{P}(\lambda) \hat{\Sigma}_s \mathbf{P}^\top(\lambda).\tag{K.5}$$

This is (49) and concludes the proof of Theorem 3.2. ■

K.1. Proof of equation (45) and (46)

Equation (45):

First note that

$$\lim_{\lambda \rightarrow 1} \hat{\Sigma}_s = \lim_{\lambda \rightarrow 1} (1-\lambda) \hat{\Sigma}_h + \lambda \hat{\Sigma}_h^{\text{d}} = \hat{\Sigma}_h^{\text{d}}\tag{K.6}$$

and

$$\begin{aligned}\lim_{\lambda \rightarrow 1} \mathbf{P}(\lambda) &= \lim_{\lambda \rightarrow 1} \left(\mathbf{S}^\top \left((1-\lambda) \hat{\Sigma}_h + \lambda \hat{\Sigma}_h^{\text{d}} \right)^{-1} \mathbf{S} \right)^{-1} \mathbf{S}^\top \left((1-\lambda) \hat{\Sigma}_h + \lambda \hat{\Sigma}_h^{\text{d}} \right)^{-1} \\ &= \left(\mathbf{S}^\top \left(\hat{\Sigma}_h^{\text{d}} \right)^{-1} \mathbf{S} \right)^{-1} \mathbf{S}^\top \left(\hat{\Sigma}_h^{\text{d}} \right)^{-1}.\end{aligned}\tag{K.7}$$

Therefore,

$$\begin{aligned}\lim_{\lambda \rightarrow 1} \mathbf{P}(\lambda) \hat{\Sigma}_s \mathbf{P}^\top(\lambda) &= \left(\mathbf{S}^\top \left(\hat{\Sigma}_h^{\text{d}} \right)^{-1} \mathbf{S} \right)^{-1} \mathbf{S}^\top \left(\hat{\Sigma}_h^{\text{d}} \right)^{-1} \hat{\Sigma}_h^{\text{d}} \left(\hat{\Sigma}_h^{\text{d}} \right)^{-1} \mathbf{S} \left(\mathbf{S}^\top \left(\hat{\Sigma}_h^{\text{d}} \right)^{-1} \mathbf{S} \right)^{-1} \\ &= \left(\mathbf{S}^\top \left(\hat{\Sigma}_h^{\text{d}} \right)^{-1} \mathbf{S} \right)^{-1} \\ &= \left(\mathbf{S}_{\text{top}}^\top \left(\hat{\Sigma}_{h,\text{top}}^{\text{d}} \right)^{-1} \mathbf{S}_{\text{top}} + \left(\hat{\Sigma}_{h,\text{bot}}^{\text{d}} \right)^{-1} \right)^{-1}.\end{aligned}\tag{K.8}$$

This is (45) and concludes the first part of the proof. ■

Equation (46):

We start by rewriting (43)

$$\begin{aligned}\hat{\Sigma}_{r,MAP} &= \frac{T/(1-\lambda)}{T+n-m} \left(P(\lambda) \hat{\Sigma}_s P^\top(\lambda) - \lambda \left(\left(\hat{\Sigma}_{h,\text{bot}}^{\text{d}} \right)^{-1} + S_{\text{top}}^\top \left(\hat{\Sigma}_{h,\text{top}}^{\text{d}} \right)^{-1} S_{\text{top}} \right)^{-1} \right) \\ &= \frac{T}{T+n-m} \left(P(\lambda) \hat{\Sigma}_h P^\top(\lambda) + \frac{\lambda}{1-\lambda} \left(P(\lambda) \hat{\Sigma}_h^{\text{d}} P^\top(\lambda) - P(1) \hat{\Sigma}_h^{\text{d}} P^\top(1) \right) \right).\end{aligned}\quad (\text{K.9})$$

We need to show that

$$\lim_{\lambda \rightarrow 1} \frac{\lambda}{1-\lambda} \left(P(\lambda) \hat{\Sigma}_h^{\text{d}} P^\top(\lambda) - P(1) \hat{\Sigma}_h^{\text{d}} P^\top(1) \right) = \mathbf{0}.\quad (\text{K.10})$$

Starting with the second term and using the form introduced in Appendix D

$$P(1) = J - J \hat{\Sigma}_h^{\text{d}} U (U^\top \hat{\Sigma}_h^{\text{d}} U)^{-1} U^\top.\quad (\text{K.11})$$

We will use the short-hand notation

$$\begin{aligned}A &= U^\top \hat{\Sigma}_h^{\text{d}} U \\ A_s &= U^\top \hat{\Sigma}_s U\end{aligned}\quad (\text{K.12})$$

and note that A_s is a function of λ . $P(1) \hat{\Sigma}_h^{\text{d}} P^\top(1)$ can be written as

$$\begin{aligned}P(1) \hat{\Sigma}_h^{\text{d}} P^\top(1) &= J \hat{\Sigma}_h^{\text{d}} J^\top + J \hat{\Sigma}_h^{\text{d}} U A^{-1} A A^{-1} U^\top \hat{\Sigma}_h^{\text{d}} J^\top - \\ &\quad J \hat{\Sigma}_h^{\text{d}} U A^{-1} U^\top \hat{\Sigma}_h^{\text{d}} J^\top - J \hat{\Sigma}_h^{\text{d}} U A^{-1} U^\top \hat{\Sigma}_h^{\text{d}} J^\top \\ &= J \hat{\Sigma}_h^{\text{d}} J^\top - J \hat{\Sigma}_h^{\text{d}} U A^{-1} U^\top \hat{\Sigma}_h^{\text{d}} J^\top.\end{aligned}\quad (\text{K.13})$$

For the first term of (K.10), note that

$$P(\lambda) = J - J \hat{\Sigma}_s U A_s^{-1} U^\top\quad (\text{K.14})$$

and, hence,

$$\begin{aligned}P(\lambda) \hat{\Sigma}_h^{\text{d}} P^\top(\lambda) &= J \hat{\Sigma}_h^{\text{d}} J^\top + J \hat{\Sigma}_s U A_s^{-1} A A_s^{-1} U^\top \hat{\Sigma}_s J^\top - J \hat{\Sigma}_s U A_s^{-1} U^\top \hat{\Sigma}_h^{\text{d}} J^\top - \\ &\quad J \hat{\Sigma}_h^{\text{d}} U A_s^{-1} U^\top \hat{\Sigma}_s J^\top.\end{aligned}\quad (\text{K.15})$$

Replacing $\hat{\Sigma}_s$ by $(1 - \lambda)\hat{\Sigma}_h + \lambda\hat{\Sigma}_h^d$ in the last two terms we get

$$\begin{aligned}
P(\lambda)\hat{\Sigma}_h^d P^\top(\lambda) &= J\hat{\Sigma}_h^d J^\top + J\hat{\Sigma}_s U A_s^{-1} A A_s^{-1} U^\top \hat{\Sigma}_s J^\top - \\
&\quad J((1 - \lambda)\hat{\Sigma}_h + \lambda\hat{\Sigma}_h^d) U A_s^{-1} U^\top \hat{\Sigma}_h^d J^\top - \\
&\quad J\hat{\Sigma}_h^d U A_s^{-1} U^\top ((1 - \lambda)\hat{\Sigma}_h + \lambda\hat{\Sigma}_h^d) J^\top \\
&= J\hat{\Sigma}_h^d J^\top + J\hat{\Sigma}_s U A_s^{-1} A A_s^{-1} U^\top \hat{\Sigma}_s J^\top - \\
&\quad (1 - \lambda)J\hat{\Sigma}_h U A_s^{-1} U^\top \hat{\Sigma}_h^d J^\top - (1 - \lambda)J\hat{\Sigma}_h^d U A_s^{-1} U^\top \hat{\Sigma}_h J^\top - \\
&\quad 2\lambda J\hat{\Sigma}_h^d U A_s^{-1} U^\top \hat{\Sigma}_h^d J^\top.
\end{aligned} \tag{K.16}$$

The second term can be written as

$$\begin{aligned}
&J\hat{\Sigma}_s U A_s^{-1} A A_s^{-1} U^\top \hat{\Sigma}_s J^\top \\
&= J((1 - \lambda)\hat{\Sigma}_h + \lambda\hat{\Sigma}_h^d) U A_s^{-1} A A_s^{-1} U^\top ((1 - \lambda)\hat{\Sigma}_h + \lambda\hat{\Sigma}_h^d) J^\top \\
&= (1 - \lambda)^2 J\hat{\Sigma}_h U A_s^{-1} A A_s^{-1} U^\top \hat{\Sigma}_h J^\top + (1 - \lambda)\lambda J\hat{\Sigma}_h U A_s^{-1} A A_s^{-1} U^\top \hat{\Sigma}_h^d J^\top + \\
&\quad (1 - \lambda)\lambda J\hat{\Sigma}_h^d U A_s^{-1} A A_s^{-1} U^\top \hat{\Sigma}_h J^\top + \lambda^2 J\hat{\Sigma}_h^d U A_s^{-1} A A_s^{-1} U^\top \hat{\Sigma}_h^d J^\top.
\end{aligned} \tag{K.17}$$

Collecting the terms we get

$$\begin{aligned}
\frac{\lambda P(\lambda)\hat{\Sigma}_h^d P^\top(\lambda)}{1 - \lambda} &= \frac{\lambda}{1 - \lambda} J\hat{\Sigma}_h^d J^\top + \lambda(1 - \lambda)J\hat{\Sigma}_h U A_s^{-1} A A_s^{-1} U^\top \hat{\Sigma}_h J^\top + \\
&\quad \lambda J\hat{\Sigma}_h U A_s^{-1} (\lambda A A_s^{-1} - I) U^\top \hat{\Sigma}_h^d J^\top + \\
&\quad \lambda J\hat{\Sigma}_h^d U A_s^{-1} (\lambda A A_s^{-1} - I) U^\top \hat{\Sigma}_h J^\top + \\
&\quad \frac{\lambda}{1 - \lambda} J\hat{\Sigma}_h^d U A_s^{-1} (\lambda A A_s^{-1} - I) U^\top \hat{\Sigma}_h^d J^\top - \\
&\quad \frac{\lambda^2}{1 - \lambda} J\hat{\Sigma}_h^d U A_s^{-1} U^\top \hat{\Sigma}_h^d J^\top.
\end{aligned} \tag{K.18}$$

In the limit $\lambda \rightarrow 1$, the second term is clearly zero. For the third through sixth terms we rewrite A_s^{-1} (using Petersen and Pedersen, 2006, eq. (157)) as

$$\begin{aligned}
A_s^{-1} &= (U^\top ((1 - \lambda)\hat{\Sigma}_h + \lambda\hat{\Sigma}_h^d) U)^{-1} \\
&= \frac{1}{\lambda} A^{-1} - \frac{1}{\lambda^2} A^{-1} U^\top \left(\frac{\hat{\Sigma}_h^{-1}}{1 - \lambda} + \frac{U A^{-1} U^\top}{\lambda} \right)^{-1} U A^{-1} \\
&= \frac{1}{\lambda} A^{-1} - \frac{1 - \lambda}{\lambda} A^{-1} U^\top (\lambda \hat{\Sigma}_h^{-1} + (1 - \lambda) U A^{-1} U^\top)^{-1} U A^{-1}.
\end{aligned} \tag{K.19}$$

Inserting (K.19) in (K.18) shows that term three and four are proportional to $1 - \lambda$ and, thus, will be zero in the limit $\lambda \rightarrow 1$. The fifth term becomes

$$\begin{aligned} \text{term5} &= \frac{\lambda}{1 - \lambda} \mathbf{J} \hat{\Sigma}_h^d \mathbf{U} \mathbf{A}_s^{-1} (\lambda \mathbf{A} \mathbf{A}_s^{-1} - \mathbf{I}) \mathbf{U}^\top \hat{\Sigma}_h^d \mathbf{J}^\top \\ &= -\lambda \mathbf{J} \hat{\Sigma}_h^d \mathbf{U} \mathbf{A}_s^{-1} (\mathbf{U}^\top (\lambda \hat{\Sigma}_h^{-1} + (1 - \lambda) \mathbf{U} \mathbf{A}^{-1} \mathbf{U}^\top)^{-1} \mathbf{U} \mathbf{A}^{-1}) \mathbf{U}^\top \hat{\Sigma}_h^d \mathbf{J}^\top. \end{aligned} \quad (\text{K.20})$$

In the limit $\lambda \rightarrow 1$, we get

$$\begin{aligned} \lim_{\lambda \rightarrow 1} \text{term5} &= -\mathbf{J} \hat{\Sigma}_h^d \mathbf{U} \mathbf{A}^{-1} (\mathbf{U}^\top (\hat{\Sigma}_h^{-1})^{-1} \mathbf{U} \mathbf{A}^{-1}) \mathbf{U}^\top \hat{\Sigma}_h^d \mathbf{J}^\top \\ &= -\mathbf{J} \hat{\Sigma}_h^d \mathbf{U} \mathbf{A}^{-1} \mathbf{U}^\top \hat{\Sigma}_h \mathbf{U} \mathbf{A}^{-1} \mathbf{U}^\top \hat{\Sigma}_h^d \mathbf{J}^\top. \end{aligned} \quad (\text{K.21})$$

As a result, we can write (K.10) as

$$\begin{aligned} &\lim_{\lambda \rightarrow 1} \frac{\lambda (\mathbf{P}(\lambda) \hat{\Sigma}_h^d \mathbf{P}^\top(\lambda) - \mathbf{P}(1) \hat{\Sigma}_h^d \mathbf{P}^\top(1))}{1 - \lambda} \\ &= -\mathbf{J} \hat{\Sigma}_h^d \mathbf{U} \mathbf{A}^{-1} \mathbf{U}^\top \hat{\Sigma}_h \mathbf{U} \mathbf{A}^{-1} \mathbf{U}^\top \hat{\Sigma}_h^d \mathbf{J}^\top - \\ &\quad \lim_{\lambda \rightarrow 1} \frac{\lambda}{1 - \lambda} \mathbf{J} \hat{\Sigma}_h^d \mathbf{U} \left(\lambda (\mathbf{U}^\top \hat{\Sigma}_s \mathbf{U})^{-1} - (\mathbf{U}^\top \hat{\Sigma}_h^d \mathbf{U})^{-1} \right) \mathbf{U}^\top \hat{\Sigma}_h^d \mathbf{J}^\top \\ &= -\mathbf{J} \hat{\Sigma}_h^d \mathbf{U} \mathbf{A}^{-1} \mathbf{U}^\top \hat{\Sigma}_h \mathbf{U} \mathbf{A}^{-1} \mathbf{U}^\top \hat{\Sigma}_h^d \mathbf{J}^\top + \\ &\quad \lim_{\lambda \rightarrow 1} \frac{\lambda}{1 - \lambda} \mathbf{J} \hat{\Sigma}_h^d \mathbf{U} \left((1 - \lambda) \mathbf{A}^{-1} \mathbf{U}^\top (\lambda \hat{\Sigma}_h^{-1} + (1 - \lambda) \mathbf{U} \mathbf{A}^{-1} \mathbf{U}^\top)^{-1} \mathbf{U} \mathbf{A}^{-1} \right) \times \\ &\quad \mathbf{U}^\top \hat{\Sigma}_h^d \mathbf{J}^\top \\ &= -\mathbf{J} \hat{\Sigma}_h^d \mathbf{U} \mathbf{A}^{-1} \mathbf{U}^\top \hat{\Sigma}_h \mathbf{U} \mathbf{A}^{-1} \mathbf{U}^\top \hat{\Sigma}_h^d \mathbf{J}^\top + \\ &\quad \mathbf{J} \hat{\Sigma}_h^d \mathbf{U} \left(\mathbf{A}^{-1} \mathbf{U}^\top (\hat{\Sigma}_h^{-1})^{-1} \mathbf{U} \mathbf{A}^{-1} \right) \mathbf{U}^\top \hat{\Sigma}_h^d \mathbf{J}^\top \\ &= -\mathbf{J} \hat{\Sigma}_h^d \mathbf{U} \mathbf{A}^{-1} \mathbf{U}^\top \hat{\Sigma}_h \mathbf{U} \mathbf{A}^{-1} \mathbf{U}^\top \hat{\Sigma}_h^d \mathbf{J}^\top + \mathbf{J} \hat{\Sigma}_h^d \mathbf{U} \mathbf{A}^{-1} \mathbf{U}^\top \hat{\Sigma}_h \mathbf{U} \mathbf{A}^{-1} \mathbf{U}^\top \hat{\Sigma}_h^d \mathbf{J}^\top \\ &= 0. \end{aligned} \quad (\text{K.22})$$

In summary, we can write the result as

$$\begin{aligned} \lim_{\lambda \rightarrow 1} \hat{\Sigma}_{r,MAP} &= \frac{T}{T + n - m} \mathbf{P}(1) \hat{\Sigma}_h \mathbf{P}^\top(1) \\ &= \frac{T}{T + n - m} (\mathbf{S}^\top (\hat{\Sigma}_h^d)^{-1} \mathbf{S})^{-1} \mathbf{S}^\top (\hat{\Sigma}_h^d)^{-1} \hat{\Sigma}_h (\hat{\Sigma}_h^d)^{-1} \mathbf{S} (\mathbf{S}^\top (\hat{\Sigma}_h^d)^{-1} \mathbf{S})^{-1}. \end{aligned} \quad (\text{K.23})$$

Since $\hat{\Sigma}_h = \left(\hat{\Sigma}_h^d \right)^{1/2} \mathbf{R}_h \left(\hat{\Sigma}_h^d \right)^{1/2}$ we have shown (46). ■

L. Proof of Corollary 3.2

When adding a REML correction term (the Hessian wrt. β of the MAP-objective) to (K.1), we get (using $e_{\text{bot}} = \mathbf{y} - \hat{\mathbf{y}}_{\text{bot}}$)

$$\begin{aligned} l_{\text{MAP,RE}}(\Sigma_r) \propto & -\frac{1}{2}(e_{\text{bot}} - \mathbf{X}\beta_{\text{top}})^\top (\mathbf{I}_T \otimes \Sigma_r^{-1}) (e_{\text{bot}} - \mathbf{X}\beta_{\text{top}}) - \\ & \frac{1}{2}(\beta_{\text{top}} - \beta_{0,\text{top}})^\top \left(\Sigma_r^{-1} \otimes \Sigma_{0,\beta}^{-1} \right) (\beta_{\text{top}} - \beta_{0,\text{top}}) - \\ & \frac{T+n+v+1}{2} \log |\Sigma_r| - \frac{1}{2} \text{Tr}(\Psi \Sigma_r^{-1}) + \\ & \frac{1}{2} \log |\mathbf{X}^\top (\mathbf{I}_T \otimes \Sigma_r^{-1}) \mathbf{X} + \Sigma_r^{-1} \otimes \Sigma_{0,\beta}^{-1}|. \end{aligned} \quad (\text{L.1})$$

We assume that $\mathbf{x}_{i,t} = \mathbf{x}_{j,t} \in \mathbb{R}^{n-m}$. In this case the determinant in the REML correction term is (see (C.6))

$$|\mathbf{X}^\top (\mathbf{I}_T \otimes \Sigma_r^{-1}) \mathbf{X} + \Sigma_r^{-1} \otimes \Sigma_{0,\beta}^{-1}| = |\Sigma_r^{-1}|^{n-m} |\mathbf{X}_{1,\cdot}^\top \mathbf{X}_{1,\cdot} + \Sigma_{0,\beta}^{-1}|^m \quad (\text{L.2})$$

and, hence,

$$\begin{aligned} l_{\text{MAP,RE}}(\Sigma_r) \propto & -\frac{1}{2}(e_{\text{bot}} - \mathbf{X}\hat{\beta}_{\text{top}})^\top (\mathbf{I}_T \otimes \Sigma_r^{-1}) (e_{\text{bot}} - \mathbf{X}\hat{\beta}_{\text{top}}) - \\ & \frac{1}{2}(\beta_{\text{top}} - \beta_{0,\text{top}})^\top \left(\Sigma_r^{-1} \otimes \Sigma_{0,\beta}^{-1} \right) (\beta_{\text{top}} - \beta_{0,\text{top}}) - \\ & \frac{T+m+v+1}{2} \log |\Sigma_r| - \frac{1}{2} \text{tr}(\Psi \Sigma_r^{-1}). \end{aligned} \quad (\text{L.3})$$

This gives the estimate (see Section K)

$$\begin{aligned} \hat{\Sigma}_{r,\text{sREML}} = & \frac{T/(1-\lambda)}{T+m+v+1} \left(\mathbf{P}(\lambda) \hat{\Sigma}_s \mathbf{P}^\top(\lambda) - \right. \\ & \left. \lambda \left(\left(\hat{\Sigma}_{h,\text{bot}}^{\text{d}} \right)^{-1} + \mathbf{S}_{\text{top}}^\top \left(\hat{\Sigma}_{h,\text{top}}^{\text{d}} \right)^{-1} \mathbf{S}_{\text{top}} \right)^{-1} + \frac{1-\lambda}{T} \Psi \right). \end{aligned} \quad (\text{L.4})$$

Inserting the values of v and Ψ we get

$$\begin{aligned} \hat{\Sigma}_{r,\text{sREML}} = & \frac{T/(1-\lambda)}{T+m+v+1} \mathbf{P}(\lambda) \hat{\Sigma}_s \mathbf{P}^\top(\lambda) \\ = & \frac{T}{T-(n-m)(1-\lambda)} \mathbf{P}(\lambda) \hat{\Sigma}_s \mathbf{P}^\top(\lambda). \end{aligned} \quad (\text{L.5})$$

This is (49) and concludes the proof of Corollary 3.2. ■

M. Additional plots for the case study

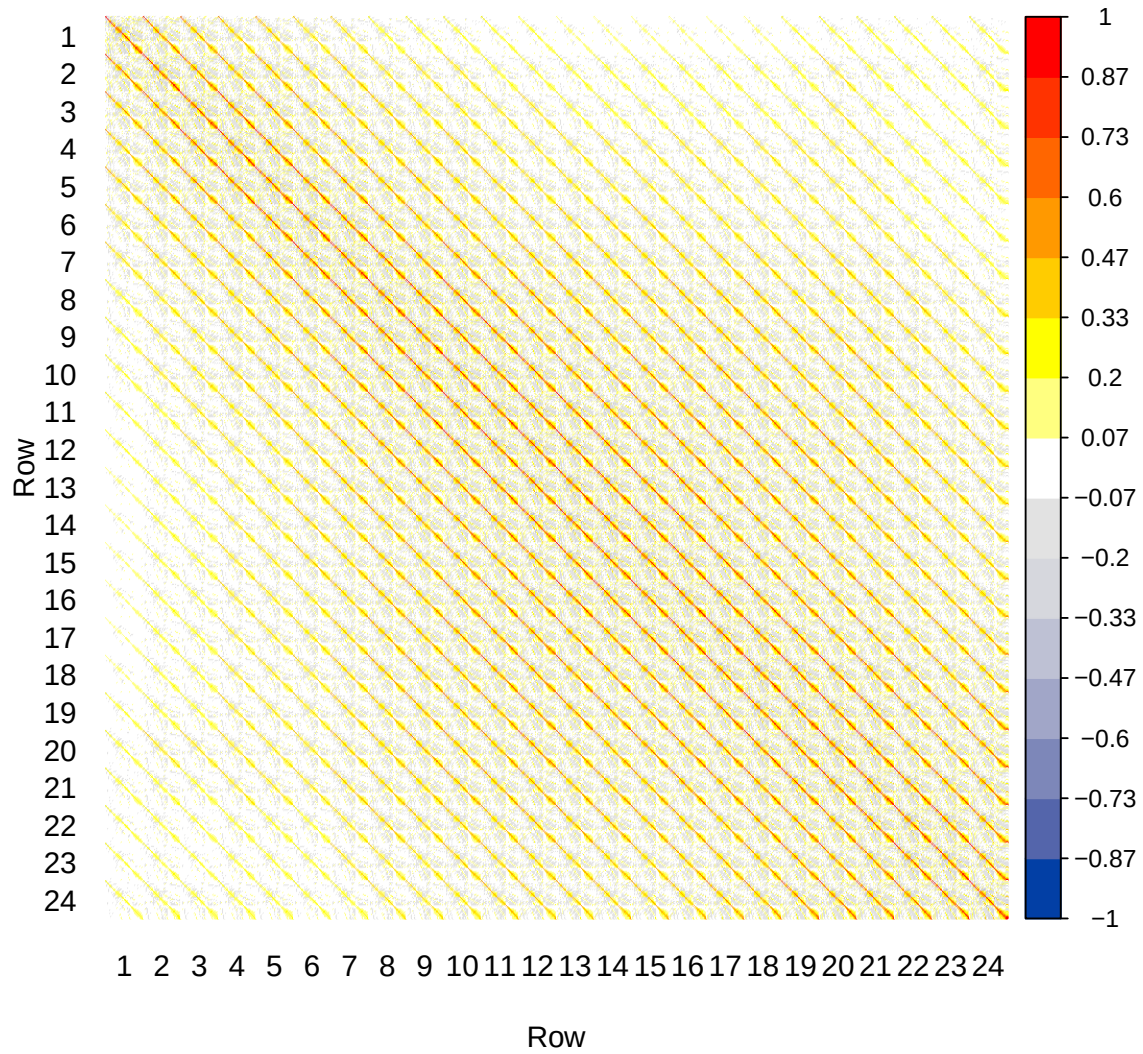


Figure M.5: Parameter correlation of all parameters in the weight matrix for area SE. The correlation patterns within rows (diagonal blocks) are constant while the between rows correlation matrices (off-diagonal blocks) display decreasing correlation with distance so that, e.g., row 1 and row 24 are weakly correlated while the correlation between row n and row $n - 1$ is strong. The within-row correlation is more clear in Figures 4 and M.8.

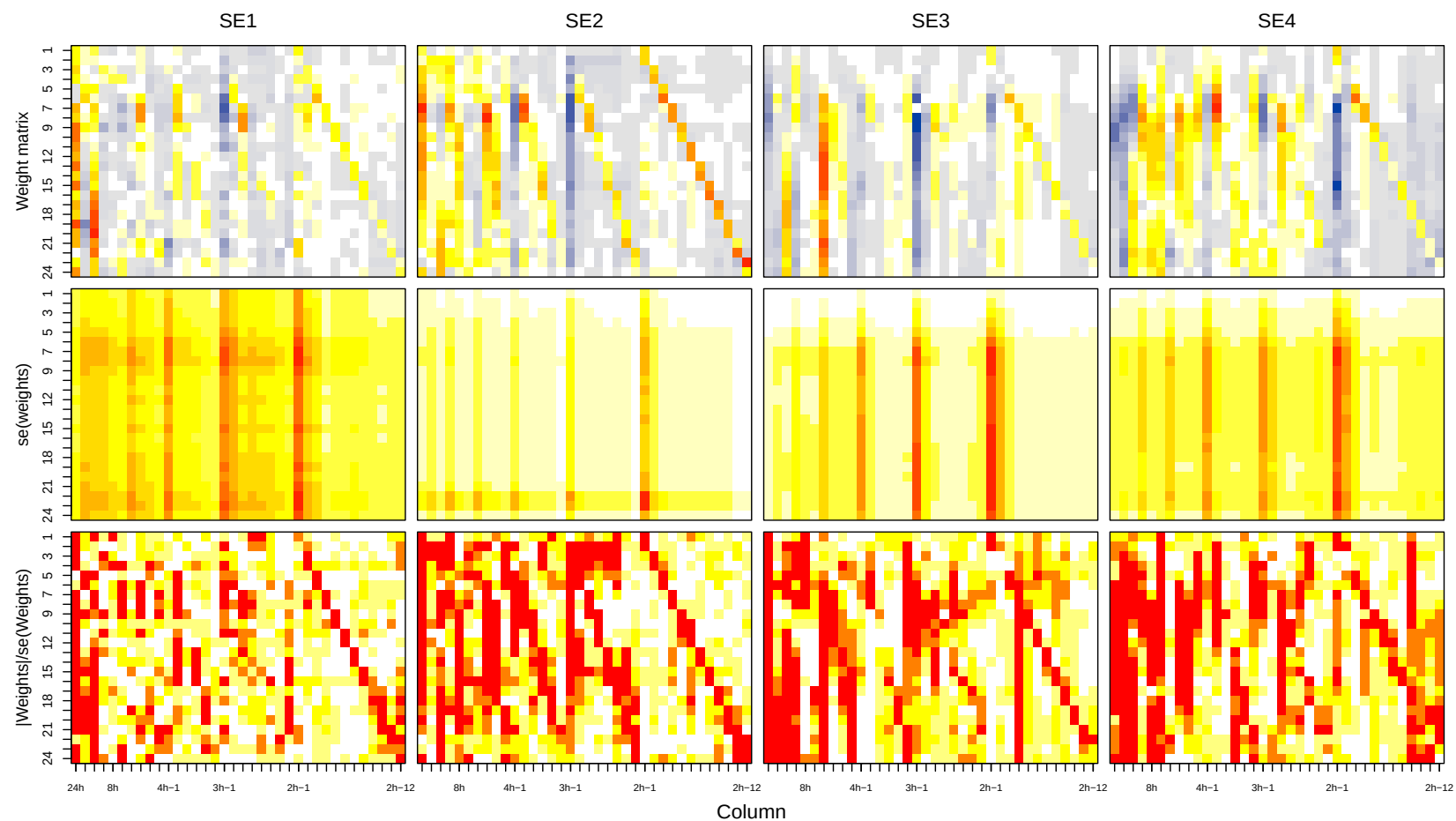


Figure M.6: SE1-SE4: Volume-weighted weight matrices, associated standard errors, and absolute value of Wald test statics. Blue colours represent negative values and red/yellow represent positive values. For the last row the colour coding is the same as in Figure 3.

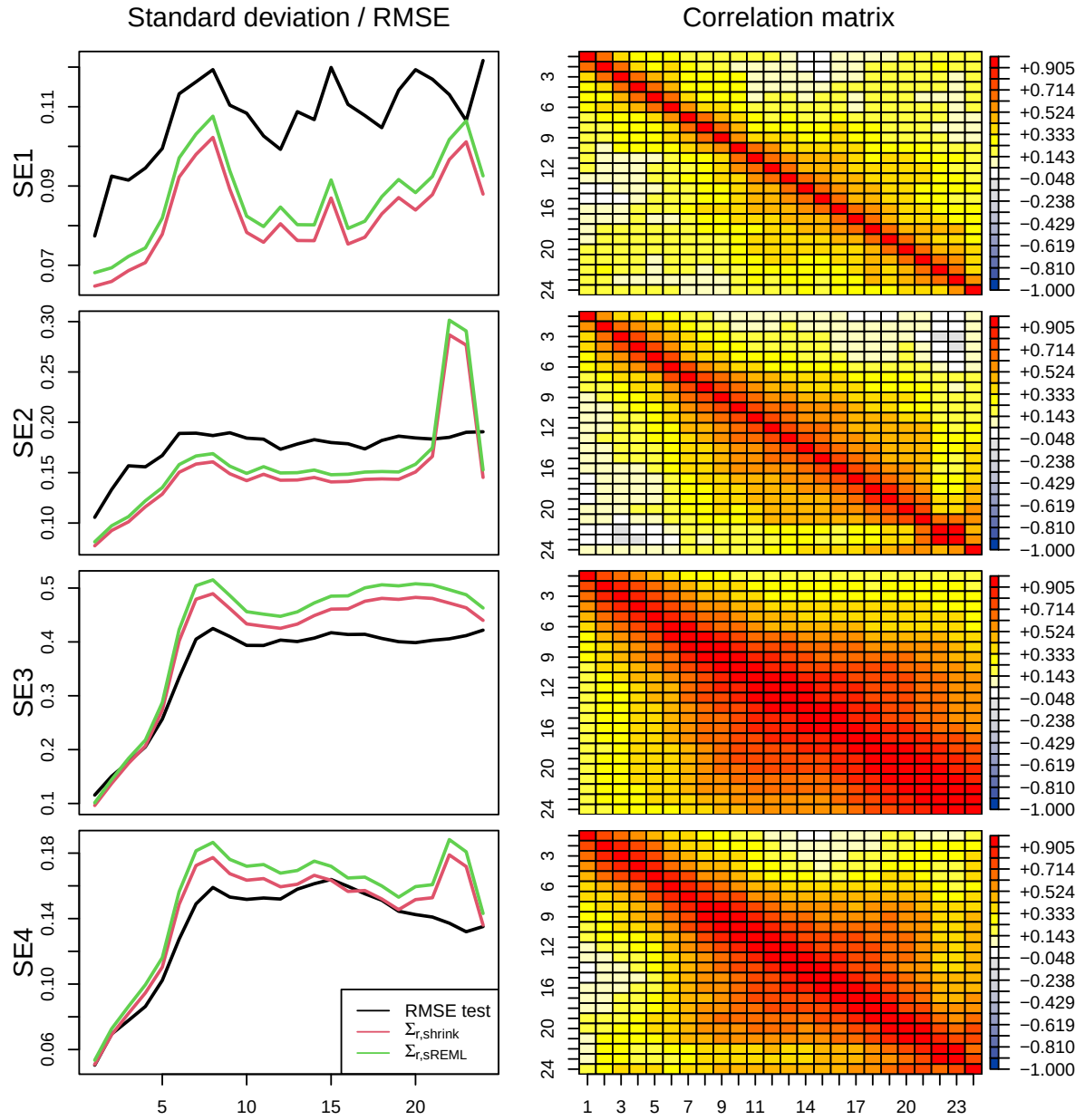


Figure M.7: Standard deviation and correlation matrix corresponding to Σ_r for areas SE1-SE4.

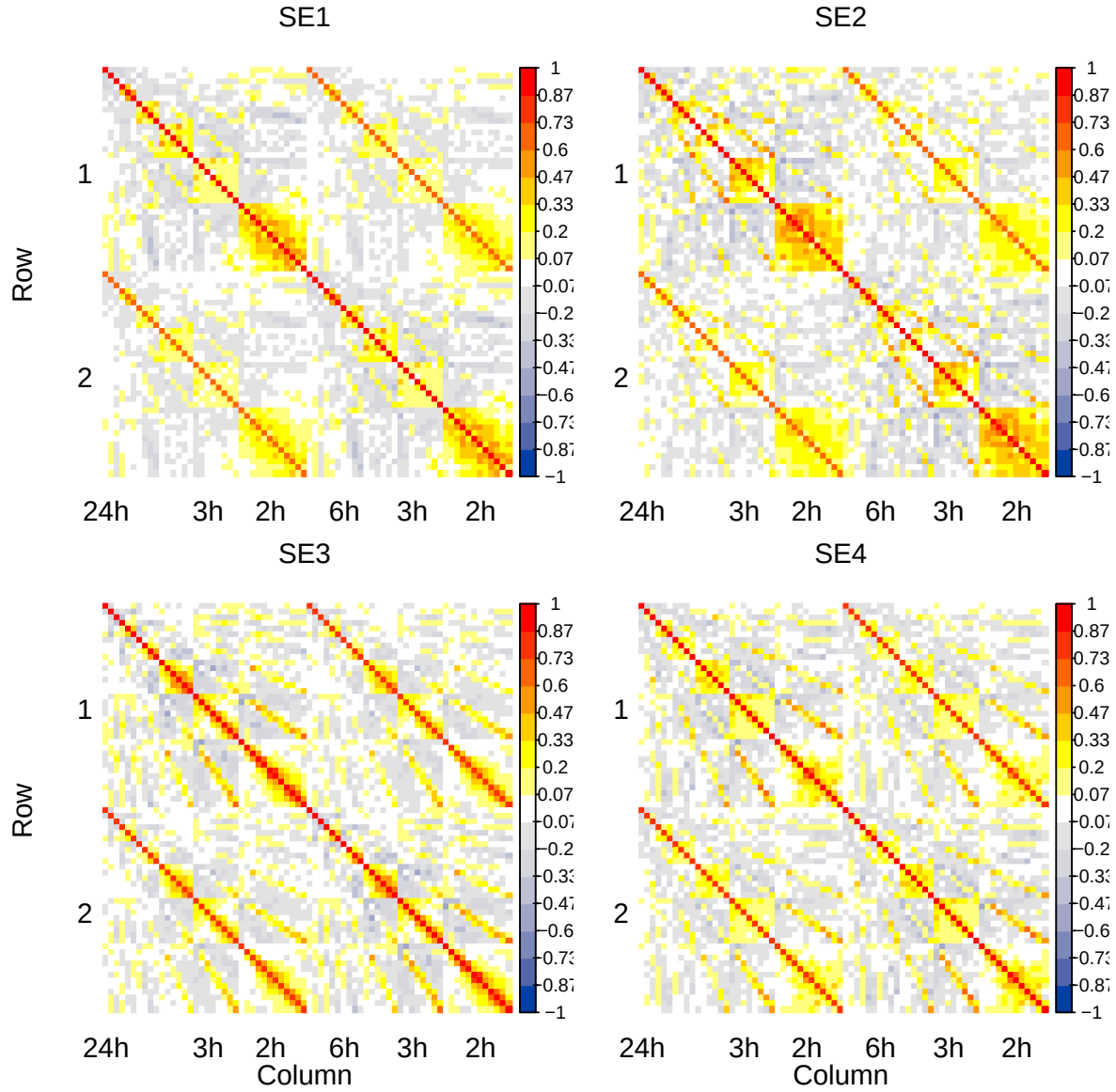


Figure M.8: Parameter correlation for rows 1–2 of the weight matrix (P_{top}) for areas SE1–SE4. The correlation between levels of the forecasts (e.g., 2h and 3h.) is generally weak as is the correlation between forecast far away in time (e.g., 2h-1 and 2h-12 for row 1). In contrast, the correlation between the same weight in different rows is generally strong.

N. Variance separation tables

	$\frac{\ \mathbf{y}_I - \hat{\mathbf{y}}_I\ ^2}{T}$	$\frac{\ \mathbf{y}_I - \tilde{\mathbf{y}}_I\ ^2}{T}$	$\frac{\ \tilde{\mathbf{y}}_I - \hat{\mathbf{y}}_I\ ^2}{T}$	$\frac{\frac{\sum SS_{mod}}{n-m}}{\frac{\sum SSE}{T-n+m}}$	$\frac{\ \mathbf{y}_I - \tilde{\mathbf{y}}_I^\lambda\ ^2}{T}$	$\frac{\ \tilde{\mathbf{y}}_I^\lambda - \hat{\mathbf{y}}_I\ ^2}{T}$	$\frac{2(\tilde{\mathbf{e}}_I^\lambda)^T \tilde{\mathbf{e}}_I^\lambda}{T}$
24h	1.56	1.177	0.379	2.94	1.297	0.213	0.046
12h	0.99	0.781	0.205	2.58	0.869	0.088	0.029
8h	0.79	0.620	0.171	2.89	0.691	0.069	0.030
6h	0.65	0.516	0.138	2.83	0.576	0.055	0.023
4h	0.50	0.393	0.105	3.23	0.436	0.046	0.016
3h	0.42	0.333	0.092	3.61	0.368	0.042	0.015
2h	0.33	0.249	0.078	4.55	0.275	0.042	0.010
1h	0.21	0.149	0.064	8.39	0.164	0.044	0.006
Total	5.45	4.217	1.233	2.65	4.675	0.599	0.175

Table N.4: Variance separation (30) on the training set for area SE1. See Table 1 for explanation.

	$\frac{\ \mathbf{y}_I - \hat{\mathbf{y}}_I\ ^2}{T}$	$\frac{\ \mathbf{y}_I - \tilde{\mathbf{y}}_I\ ^2}{T}$	$\frac{\ \tilde{\mathbf{y}}_I - \hat{\mathbf{y}}_I\ ^2}{T}$	$\frac{\frac{\sum SS_{mod}}{n-m}}{\frac{\sum SSE}{T-n+m}}$	$\frac{\ \mathbf{y}_I - \tilde{\mathbf{y}}_I^\lambda\ ^2}{T}$	$\frac{\ \tilde{\mathbf{y}}_I^\lambda - \hat{\mathbf{y}}_I\ ^2}{T}$	$\frac{2(\tilde{\mathbf{e}}_I^\lambda)^T \tilde{\mathbf{e}}_I^\lambda}{T}$
24h	5.80	4.412	1.391	2.88	5.073	0.597	0.132
12h	3.84	3.094	0.746	2.46	3.503	0.234	0.102
8h	3.06	2.471	0.590	2.60	2.791	0.194	0.076
6h	2.66	2.062	0.601	3.45	2.340	0.264	0.058
4h	2.13	1.579	0.552	4.85	1.788	0.289	0.053
3h	1.79	1.271	0.516	6.15	1.439	0.299	0.050
2h	1.20	0.889	0.313	5.21	1.005	0.167	0.029
1h	0.71	0.516	0.192	6.30	0.582	0.108	0.018
Total	21.19	16.292	4.901	3.07	18.522	2.153	0.518

Table N.5: Variance separation (30) on the training set for area SE2. See Table 1 for explanation.

	$\frac{\ \mathbf{y}_I - \hat{\mathbf{y}}_I\ ^2}{T}$	$\frac{\ \mathbf{y}_I - \tilde{\mathbf{y}}_I\ ^2}{T}$	$\frac{\ \tilde{\mathbf{y}}_I - \hat{\mathbf{y}}_I\ ^2}{T}$	$\frac{\frac{\sum SS_{mod}}{n-m}}{\frac{\sum SSE}{T-n+m}}$	$\frac{\ \mathbf{y}_I - \tilde{\mathbf{y}}_I^\lambda\ ^2}{T}$	$\frac{\ \tilde{\mathbf{y}}_I^\lambda - \hat{\mathbf{y}}_I\ ^2}{T}$	$\frac{2(\tilde{\mathbf{e}}_I^\lambda)^T \tilde{\mathbf{e}}_I^\lambda}{T}$
24h	110.2	49.58	60.64	11.18	61.90	44.94	3.39
12h	50.1	30.07	20.04	6.72	37.06	11.94	1.11
8h	32.3	22.28	10.02	4.53	27.41	3.87	1.03
6h	25.3	17.77	7.53	4.14	21.85	2.59	0.86
4h	17.9	12.42	5.48	4.51	15.24	2.09	0.57
3h	13.8	9.55	4.25	4.56	11.71	1.61	0.48
2h	9.1	6.53	2.60	3.94	7.99	0.82	0.31
1h	4.8	3.35	1.40	4.05	4.10	0.49	0.16
Total	263.5	151.55	111.96	4.42	187.27	68.35	7.90

Table N.6: Variance separation (30) on the training set for area SE3. See Table 1 for explanation.

	$\frac{\ \mathbf{y}_I - \hat{\mathbf{y}}_I\ ^2}{T}$	$\frac{\ \mathbf{y}_I - \tilde{\mathbf{y}}_I\ ^2}{T}$	$\frac{\ \tilde{\mathbf{y}}_I - \hat{\mathbf{y}}_I\ ^2}{T}$	$\frac{\frac{\sum SS_{mod}}{n-m}}{\frac{\sum SSE}{T-n+m}}$	$\frac{\ \mathbf{y}_I - \tilde{\mathbf{y}}_I^\lambda\ ^2}{T}$	$\frac{\ \tilde{\mathbf{y}}_I^\lambda - \hat{\mathbf{y}}_I\ ^2}{T}$	$\frac{2(\tilde{\mathbf{e}}_I^\lambda)^T \tilde{\mathbf{e}}_I^\lambda}{T}$
24h	9.92	5.660	4.265	6.89	6.760	2.781	0.384
12h	5.46	3.455	2.002	5.69	4.106	1.158	0.193
8h	3.79	2.649	1.140	4.28	3.129	0.544	0.116
6h	3.08	2.179	0.898	4.19	2.570	0.403	0.104
4h	2.23	1.544	0.687	4.79	1.821	0.338	0.072
3h	1.78	1.211	0.569	5.22	1.425	0.292	0.063
2h	1.19	0.831	0.359	4.80	0.979	0.165	0.046
1h	0.67	0.439	0.227	5.68	0.516	0.128	0.022
Total	28.11	17.968	10.147	4.24	21.306	5.809	1.000

Table N.7: Variance separation (30) on the training set for area SE4. See Table 1 for explanation.