Enhancing Forecast Reconciliation: A Study of Alternative Covariance Estimators

Vincent Su Shanika Wickramasuriya (supv.)
George Athanasopoulos (supv.)

Abstract

A collection of time series connected via a set of linear constraints is known as hierarchical time series. Forecasting these series without respecting the hierarchical nature of the data can lead to incoherent forecasts across aggregation levels and lower accuracy. To mitigate this issue, various forecast reconciliation approaches have been proposed in the literature, where the individual forecasts are adjusted to satisfy the aggregation constraints. Among these, **MinT** (Minimum Trace) is widely used, however, it requires a good estimate of the covariance matrix of the base forecast errors. The current practice is to use the shrinkage estimator (often shrinking toward a diagonal matrix), but it lacks flexibility and might not fully utilise the prominent latent structure presented. In this project, we aim to assess the forecasting performance of MinT when different covariance estimators are used, namely NOVELIST (NOVEL Integration of the Sample and Thresholded Covariance), PC-adjusted estimators (taking the latent factors directly into its construction), and others.

1 Introduction

In time series forecasting, aggregation occurs in a variety of settings. For example, Starbucks operates in many countries, and each country has multiple cities where they have outlets. The sales data is structured *hierarchically*: the top level is the total sales across all countries, which are the sum of sales in each country, which in turn are the sum of sales in each city, and finally the sum of sales from each outlet in the city. As a result, there are over 50,000 sales series across all *aggregation levels*, and decision makers need forecasts at each level to manage inventory and plan marketing strategies effectively. The hierarchy can be even more complex if we consider the sales of different kinds of drinks (e.g., coffees, teas, refreshers) at each aggregation level. Such hierarchical structures also arises in many other decision-making contexts, from supply chains (Angam et al., 2025; Seaman & Bowman, 2022) and energy

planning (Di Modica et al., 2021), to macroeconomics (El Gemayel et al., 2022; Li et al., 2019) and tourism analysis (Athanasopoulos et al., 2009). Stakeholders in these settings need forecasts at several aggregation levels to allocate resources and manage risk.

In practice, when forecasts are produced for all series (often called *base forecasts*), they typically violate the aggregation constraints observed in the data (the sum of all countries' sales does not equal the total sales). Such forecasts are called *incoherent*. First, they undermine downstream decisions that require internal consistency, and second, our forecast performance suffers. To overcome these issues, forecast reconciliation was introduced. Forecast reconciliation, a post-processing step, utilises the information from the hierarchical structure and data to adjust the initially produced base forecasts, so that the resulting forecasts are *coherent* (i.e., respecting the aggregation constraints). It first introduced by Hyndman et al. (2011), and later developed by Erven & Cugliari (2015), Hyndman et al. (2016), Ben Taieb & Koo (2019), Wickramasuriya et al. (2019), Wickramasuriya et al. (2020), and others to enhance point forecast performance. Athanasopoulos et al. (2024) provided a comprehensive review of the literature on forecast reconciliation. Among the modern methods, the Min Trace (MinT) approach developed by Wickramasuriya et al. (2019) provides good theoretical properties of minimising the total variance of the reconciled forecast errors, fast computational speed, and significant empirical improvements in many applications. It has become a standard method for forecast reconciliation, implemented in popular R and Python softwares/packages (Nixtla, 2025; O'Hara-Wild et al., 2024).

A difficulty with MinT is in estimating the covariance matrix of the base forecast errors, especially for 2-step-ahead forecasts and beyond. This is a high-dimensional covariance estimation problem, where the number of series n is often larger than the time dimension T. Wickramasuriya et al. (2019) proposed using a shrinkage estimator with diagonal target from Schäfer & Strimmer (2005) to estimate the 1-step-ahead covariance matrix, which is then scaled by a constant to approximate the h-step-ahead covariance matrix. While this approach is relatively simple and guarantees positive definiteness, it has three main shortcomings. First, it uniformly shrinks all off-diagonal elements towards zeros with equal weights, lacking flexibility to effectively preserve strong signals in the data. Second, in many real-world applications, hierarchical time series data often exhibit a prominent principal components structure, which is not fully utilised. Third, the proportionality relationship between the h-step-ahead and 1-step-ahead forecast error covariance matrices might not hold in practice.

The theorectical advantage of MinT might not be achieved in finite sample settings if the covariance estimate is not close enough to the true covariance matrix. Despite its importance, not many researchers have explored this issue in the current literature, except Carrara et al. (2025), who proposed a Double Shrinkage estimator for MinT. The idea is to introduce a second shrinkage target, which incorporates conditional dependence information suggested by the hierarchical structure. This paper is still in its early stage, and does not fully address the shortcomings mentioned above.

In the current literature, there are many advancements in high-dimensional covariance estimation that can be potentially useful for MinT. Building upon the shrinkage estimator, Huang

& Fryzlewicz (2019) introduces a soft-thresholded target matrix instead of the diagonal target. Extending beyond linear shrinkage, Ledoit & Wolf (2012) proposed a nonlinear shrinkage estimator that replaces each sample eigenvalue with a data-driven nonlinear transformation of itself, which is further developed by Ledoit & Wolf (2020). On the other hand, Fan et al. (2013) introduced an estimator that directly preserves the latent factors structure, decomposing the covariance matrix into a low-rank component and a sparse component, and applying thresholding to the latter. There are also approaches related to thresholding, such as the adaptive thresholding estimator for sparse covariance matrices by Cai & Liu (2011). The list of modern high-dimensional covariance estimators is long, presenting many opportunities to explore their potential in improving the performance of MinT.

The purpose of this paper is to explore the shortcomings of the shrinkage estimator and introduce alternative estimators that address these issues one by one. We evaluate the performance of MinT with these different covariance estimators in both point and probabilistic forecast reconciliation settings. In our empirical results from the complex, large-hierarchy dataset, we observe the following: (i) MinT with NOVELIST estimator with brings improvements over the shrinkage estimator under the probabilistic setting, as its predictive distributions have higher coverage rates; (ii) Covariance estimators that utilise the latent factors structure consistently produce better point and probabilistic forecasts than those that do not, when the data is evidently driven by such structure; (iii) Scaling the 1-step-ahead covariance matrix to approximate the h-step-ahead one gives significantly better performance than estimators that do not use this assumption, suggesting that the scaling assumption is not entirely incorrect. The estimators relaxing on the proportionality assumption that we propose can also be further extended to h-step-ahead probabilistic forecast reconciliation.

The remainder of the paper is organised as follows. Section 2 provides the basic theoretical framework for hierarchical time series and forecasting, and Minimum Trace approach, introducing notations, terminologies, and motivations for alternative estimators. Section 3 walks through the main covariance estimators this paper explores, and argues their strengths and weaknesses. Section 5 covers the simulation design and currently explores the performance of NOVELIST on MinT. Section 6 shows a real-world application of MinT with NOVELIST, which produces results that did not occured in the simulation settings, suggesting further inspection and analysis.

2 Theoretical Framework

2.1 Hierarchical Time Series

Hierarchical time series are multivariate time series $\boldsymbol{y}_t \in \mathbb{R}^n$ organised in a structure where the series adheres to some *constraints*. For example, Figure 1 illustrates a simple 2-level hierarchical structure with one top-level series $y_{Tot,t}$, two middle-level series $(y_{A,t}, y_{B,t})'$, and

four bottom-level series $(y_{A1,t}, y_{A2,t}, y_{B1,t}, y_{B2,t})'$. Here, the *aggregation constraints* imply that $y_{Tot,t} = y_{A,t} + y_{B,t}$, $y_{A,t} = y_{A1,t} + y_{A2,t}$, and $y_{B,t} = y_{B1,t} + y_{B2,t}$.

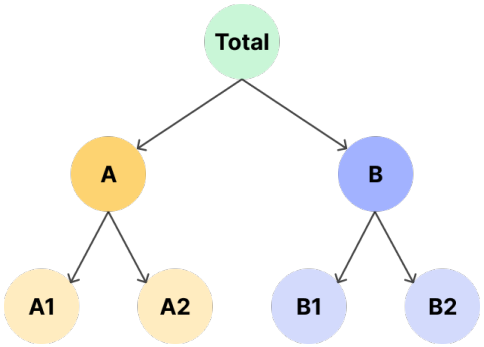

Figure 1: A 2-level hierarchical tree structure

The bottom-level (or most disaggregated) series are denoted as $\boldsymbol{b}_t \in \mathbb{R}^{n_b}$. Thus, the hierarchical time series can be represented as:

$$\boldsymbol{y}_t = \boldsymbol{S}\boldsymbol{b}_t,$$

where $\boldsymbol{S} \in \mathbb{R}^{n \times n_b}$ is a summing matrix that aggregates the bottom-level to all-level series. The summing matrix $\boldsymbol{S}$ for the tree structure in Figure 1 is:

$$\boldsymbol{S} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ & \boldsymbol{I_4} & & \end{bmatrix}.$$

The matrix $\boldsymbol{S}$ encodes the aggregation constraints presented in the structure. Hence, the columns of $\boldsymbol{S}$ span a linear subspace. Any observation $\boldsymbol{y}_t$ that lies inside this subspace is called *coherent*, while those outside are *incoherent*. We refer to the subspace spanned by $\boldsymbol{S}$ as the *coherent subspace* $\mathfrak{s} \in \mathbb{R}^{n_b}$.

This setting is not restricted to hierarchical (nested) structures. When there are attributes of interest that are crossed, such as the company sales at any aggregation level (company-wise, city-wise, or outlet-wise) is also considered by kinds of products, the structure is described as a *grouped structure*. As illustrated in Figure 2, the aggregation or disaggregation paths are not unique. These constraints formed by the grouped structure can also be represented using a summing matrix $\boldsymbol{S}$. For simplicity, we refer to both of these structures as hierarchical structure, we will distinguish between them if and when it is necessary.

When we produce forecasts for each individual series, referred to as *base forecasts* $\hat{\boldsymbol{y}}_{t+h|t}$, they often do not respect the aggregation constraints, and thus are incoherent. Coherency can

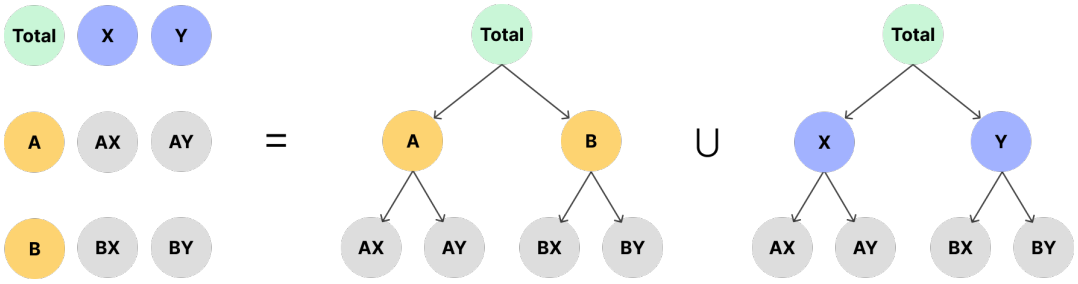

Figure 2: A 2-level grouped structure, which can be considered as the union of two hierarchical trees with common top and bottom level series

be achieved by linearly projecting the base forecasts onto the coherent subspace $\mathfrak{s}$ using a projection matrix $\boldsymbol{P}$: $\tilde{\boldsymbol{y}}_{t+h|t} = \boldsymbol{P}\hat{\boldsymbol{y}}_{t+h|t}$, where $\tilde{\boldsymbol{y}}_{t+h|t}$ are the *reconciled forecasts*. A schematic illustration of this projection is depicted in Figure 3.

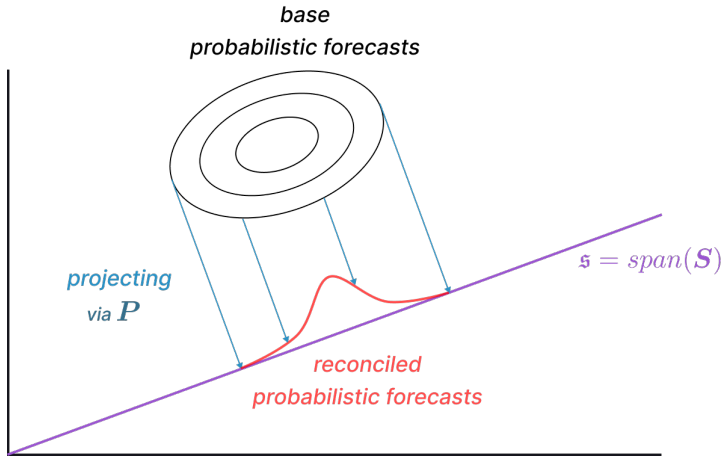

Figure 3: Geometry of probabilistic forecast reconciliation. The eliptical base forecast distribution is projected orthogonally onto the coherent subspace (purple line), resulting in the reconciled forecast distribution (red). The projection is defined by the projection matrix $\boldsymbol{P}$, and MinT allows oblique projections. Note that this figure is schematic since most applications are high-dimensional.

Many existing reconciliation methods including the OLS (Hyndman et al., 2011), WLS (Hyndman et al., 2016), and MinT (Wickramasuriya et al., 2019) express the projection matrix as $\boldsymbol{P} = \boldsymbol{SG}$, for a suitable $n_b \times n$ mapping matrix $\boldsymbol{G}$. The idea is to map the base forecasts of all levels $\hat{\boldsymbol{y}}_{t+h|t}$ down into the bottom level, which is then aggregated to the higher levels by $\boldsymbol{S}$. Since the projection matrix $\boldsymbol{P}$ is idempotent and symmetric, $\boldsymbol{G}$ must satisfy the condition $\boldsymbol{SGS} = \boldsymbol{S}$. Thus, we generally have the mapping matrix $\boldsymbol{G} = (\boldsymbol{S}'\boldsymbol{M}^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{M}^{-1}$, for some positive definite matrix $\boldsymbol{M}$ (Gamakumara, 2020).

When setting $\boldsymbol{M} = \boldsymbol{I}_n$, the identity matrix, we obtain the OLS method, which also corresponds to an orthogonal projection onto the coherent subspace (similar to Figure 3). Research have been done to introduce $\boldsymbol{M}$ that better utilise the inherent information from the observed data, to construct oblique projections and deliver better-performing forecasts.

2.2 The Minimum Trace (MinT) Reconciliation

Wickramasuriya et al. (2019) showed that by setting $\boldsymbol{M} = \boldsymbol{W}_h = \mathbb{E}(\hat{\boldsymbol{e}}_{t+h|t}\ \hat{\boldsymbol{e}}'_{t+h|t})$, the covariance matrix of the h-step-ahead base forecast errors $\hat{\boldsymbol{e}}_{t+h|t} = \boldsymbol{y}_{t+h} - \hat{\boldsymbol{y}}_{t+h|t}$, we essentially minimise the total variance of the reconciled forecast errors across all series. In which, it is minimising the trace of $\text{Var}[y_{t+h} - \tilde{y}_{t+h|t}] = \boldsymbol{S}\boldsymbol{G}_h\boldsymbol{W}_h\boldsymbol{G}'_h\boldsymbol{S}'$. This method is called Minimum Trace (MinT) reconciliation. The matrix $\boldsymbol{G}_h$ is thus given by:

$$\boldsymbol{G}_h = (\boldsymbol{S}'\boldsymbol{W}_h^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}_h^{-1}.$$

This is initially derived under point forecast reconciliation setting. However, Wickramasuriya (2024) showed that under the Gaussian base forecast distribution assumption, MinT also yields a Gaussian reconciled forecast distribution. Let h-step-ahead base forecast distribution be $\hat{\boldsymbol{y}}_{t+h|t} \sim \mathcal{N}(\hat{\boldsymbol{y}}_{t+h|t}, \boldsymbol{W}_h)$, then the reconciled forecast distribution is given by $\tilde{\boldsymbol{y}}_{t+h|t} \sim \mathcal{N}(\tilde{\boldsymbol{y}}_{t+h|t}, \boldsymbol{S}\boldsymbol{G}_h\boldsymbol{W}_h\boldsymbol{G}'_h\boldsymbol{S}')$. Additionally, MinT also minimises the logarithmic score of the reconciled predictive distribution among all projection matrices.

For probabilistic forecasts, this paper focuses on evaluating the performance of MinT with different covariance estimators under the Gaussian framework. Even so, it is worth to mention that the methods can be extended to non-Gaussian settings by bootstrapping (Gamakumara, 2020; Panagiotelis et al., 2023).

2.3 Shrinkage Estimator for MinT

The MinT solution hinges on a reliable, positive-definite estimate of $\boldsymbol{W}_h$, which comes in both the mapping matrix $\boldsymbol{G}_h$ and the reconciled forecast variance $\boldsymbol{S}\boldsymbol{G}_h\boldsymbol{W}_h\boldsymbol{G}'_h\boldsymbol{S}'$.

However, the covariance matrix $\boldsymbol{W}_h$ is often not available in theoretical derivation, and is challenging to estimate in high-dimensional setting where the number of series n is larger than the time dimension T. To tackle this issue, the original paper Wickramasuriya et al. (2019) assumed a proportionality relationship between $\hat{\boldsymbol{W}}_h^g = k_h g(\hat{\boldsymbol{W}}_1)$, where $\hat{\boldsymbol{W}}_1$ is the covariance matrix of the in-sample 1-step-ahead base residuals (to approximate $\boldsymbol{W}_1$) and k_h is a positive scaling constant (which will be cancelled out in point-forecast reconciliation). The function $g(.)$ is a covariance estimator that produces a positive-definite matrix, the main focus of this paper.

The authors suggested using the shrinkage estimator with diagonal target from Schäfer & Strimmer (2005), given by:

$$\hat{\boldsymbol{W}}_1^S = \lambda_S \operatorname{diag}(\hat{\boldsymbol{W}}_1) + (1 - \lambda_S)\hat{\boldsymbol{W}}_1 \,,$$

where $\operatorname{diag}(\hat{\boldsymbol{W}}_1)$ comprises only the diagonal elements of $\hat{\boldsymbol{W}}_1$. We refer to any $\lambda_S \in [0, 1]$ as the shrinkage intensity parameter, the subscript specifies the shrinkage estimator it belongs to. This approach shrinks the covariance matrix $\hat{\boldsymbol{W}}_1$ towards its diagonal matrix, meaning the off-diagonal elements are shrunk towards zero while the diagonal ones remain unchanged.

Schäfer & Strimmer (2005) also proposed an estimate of the optimal shrinkage intensity parameter λ_S:

$$\hat{\lambda}_S = \frac{\sum_{i \neq j} \widehat{Var}(\hat{r}_{ij})}{\sum_{i \neq j} \hat{r}_{ij}^2} \,,$$

where $\hat{r}_{ij}$ is the ijth element of $\hat{\boldsymbol{R}}_1$, the 1-step-ahead sample correlation matrix (obtained from $\hat{\boldsymbol{W}}_1$). The optimal estimate is obtained by minimising $MSE(\hat{\boldsymbol{W}}_1) = Bias(\hat{\boldsymbol{W}}_1)^2 + Var(\hat{\boldsymbol{W}}_1)$. More specifically, we trade the unbiasedness of the sample covariance matrix for a lower variance.

Despite its relative simplicity and guaranteed positive definiteness, MinT with shrinkage estimator has three main shortcomings.

Problem 1: Uniform shrinkage

The shrinkage estimator shrinks all off-diagonal elements towards zeros with equal weights λ_S. We might prefer to better preserve strong signals, and largely reduce the effects of small, noisy correlations. This is especially true in high-dimensional settings, where only a few series to be truly correlated; or due to the aggregation effects, bottom-level series will correlate more strongly with their parents. The shrinkage estimator lacks this flexibility.

Problem 2: Latent factors

The hierarchical time series data often exhibit a prominent principal components structure, which is not fully taken advantage. Taking an example of the Australian domestic overnight trips data set (*Tourism Research Australia*, 2024), where the national trips are disaggregated into states and territories, and further into regions. We then fit ETS models to all series, using the algorithm from Fabletools R package (O'Hara-Wild et al., 2024), and compute the one-step-ahead in-sample base forecast residual covariance matrix $\hat{\boldsymbol{W}}_1$. The twenty largest eigenvalues of the covariance matrix are plotted in Figure 4. We can see that the point of inflexion occurs

at the component with 5th largest eigenvalue, indicating a prominent principal components structure.

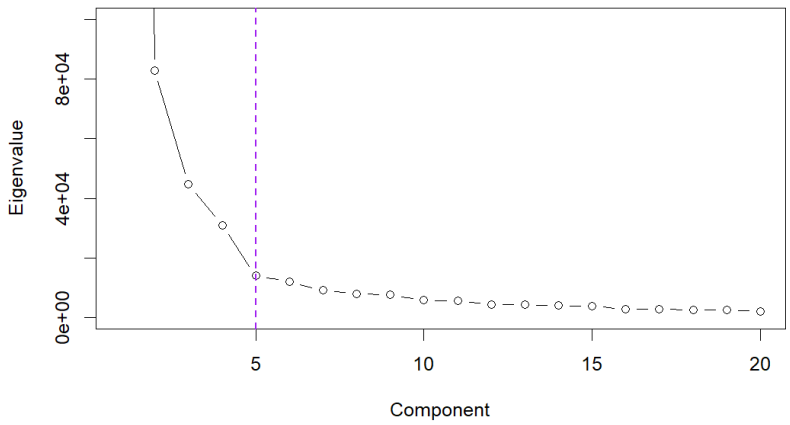

Figure 4: Twenty largest eigenvalues of one-step-ahead in-sample base forecast error covariance, Australian domestic overnight trips

Problem 3: Proportional scaling for h-step-ahead

The proportionality relationship $\hat{\boldsymbol{W}}_h^g = k_h g(\hat{\boldsymbol{W}}_1)$ might not hold in practice. The covariance structure of h-step-ahead base forecast errors can be different from that of 1-step-ahead ones.

In the next sections, we will explore several options that tackle these issues one by one.

3 Covariance Estimation Approaches

3.1 NOVELIST Estimator

The NOVELIST (NOVEL Integration of the Sample and Thresholded Covariance) estimator, proposed by Huang & Fryzlewicz (2019), introduces a way to control the target matrix's sparsity, retaining strong correlations while discarding weak, noisy effects. NOVELIST offers more flexibility than the shrinkage estimator, which is useful when we believe that only a few variables are truly correlated.

The method is based on the idea of soft-thresholding the sample covariance matrix, then performing shrinkage towards this thresholded version. This introduces an extra parameter,

the threshold δ, which is used to control the amount of soft-thresholding. The NOVELIST estimator is given by:

$$\hat{\boldsymbol{W}}_1^N = \lambda_\delta \hat{\boldsymbol{W}}_{1,\delta} + (1 - \lambda_\delta)\hat{\boldsymbol{W}}_1, \tag{1}$$

where $\hat{\boldsymbol{W}}_{1,\delta}$ is the thresholded version of $\hat{\boldsymbol{W}}_1$. By convenient setting, we can rewrite it in terms of correlation:

$$\hat{\boldsymbol{R}}_1^N = \lambda_\delta \hat{\boldsymbol{R}}_{1,\delta} + (1 - \lambda_\delta)\hat{\boldsymbol{R}}_1, \tag{2}$$

where, $\hat{\boldsymbol{R}}_{1,\delta}$ is the thresholded correlation matrix, where each element is regularised by:

$$\hat{r}_{1,ij}^{\delta} = \text{sign}(\hat{r}_{1,ij}) \max(|\hat{r}_{1,ij}| - \delta,\ 0), \tag{3}$$

where $\delta \in [0, 1]$ is the threshold parameter. For a given threshold δ, Huang & Fryzlewicz (2019) derived an analytical expression for the optimal shrinkage intensity parameter $\lambda(\delta)$ using Ledoit-Wolf's lemma [Ledoit & Wolf (2003)], following similar logic to Schäfer & Strimmer (2005). It can be computed as:

$$\hat{\lambda}(\delta) = \frac{\sum_{i \neq j} \widehat{Var}(\hat{r}_{1,ij})\ \mathbf{1}(|\hat{r}_{1,ij}| \leq \delta)}{\sum_{i \neq j} (\hat{r}_{1,ij} - \hat{r}_{1,ij}^{\delta})^2}, \tag{4}$$

where $\mathbf{1}(.)$ is the indicator function.

On the other hand, the optimal threshold $\hat{\delta}$ does not have a closed-form solution, and is typically obtained by executing a rolling-window cross-validation procedure. The idea is to find the threshold δ^*, with the corresponding λ^* and $\hat{\boldsymbol{R}}_1^N(\delta^*, \lambda^*)$, that minimises the average out-of-sample 1-step-ahead reconciled forecast mean squared error over all windows. The formal algorithm is given in Section 3.1.1.

We also tested out minimising 1- to h-step-ahead overall MSE in the cross-validation procedure. Surprisingly, it returns almost the same best threshold parameter is in the 1-step-ahead case above. Note that when $\delta \in [\max_{i \neq j}|\hat{r}_{1,ij}|,\ 1]$, the NOVELIST estimator collapses to the shrinkage estimator, and when $\delta = 0$, it becomes the sample covariance matrix. An additional concern is that the estimator does not guarantee to be positive definite, but we can use Higham (2002) algorithm to compute the nearest positive definite matrix if needed.

3.1.1 NOVELIST cross-validation algorithm

It is only required to fit the base models once on the whole training data $\{\boldsymbol{y}_t\}_{t=1}^T$, and obtain the in-sample fitted values $\{\hat{\boldsymbol{y}}_t\}_{t=1}^T$.

Algorithm 1 Cross-validation procedure

Input: Observations and fitted values $\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t \in \mathbb{R}^n$ for $t = 1, \dots, T$, set of threshold candidates Δ, window size v.
$\hat{\boldsymbol{e}}_t = \boldsymbol{y}_t - \hat{\boldsymbol{y}}_t$ for $t = 1, \dots, T$
for $i = v : T-1$ **do**
    $j = i - v + 1$
    $\hat{\boldsymbol{W}}_j = \frac{1}{v}\sum_{t=j}^{i} \hat{\boldsymbol{e}}_t \hat{\boldsymbol{e}}_t'$
    $\hat{\boldsymbol{D}}_j = \text{diag}(\hat{\boldsymbol{W}}_j)$
    $\hat{\boldsymbol{R}}_j = \hat{\boldsymbol{D}}_j^{-1/2} \hat{\boldsymbol{W}}_j \hat{\boldsymbol{D}}_j^{-1/2}$
    for $\delta \in \Delta$ **do**
        Compute thresholded correlation $\hat{\boldsymbol{R}}_{j,\delta}$ using Equation 5
        Compute $\hat{\lambda}_{j,\delta}$ using Equation 6
        Compute $\hat{\boldsymbol{R}}_{j,\delta}^N$ using Equation 4
        $\hat{\boldsymbol{W}}_{j,\delta}^N = \hat{\boldsymbol{D}}_j^{1/2} \hat{\boldsymbol{R}}_{j,\delta}^N \hat{\boldsymbol{D}}_j^{1/2}$
        $\boldsymbol{G} = (\boldsymbol{S}' \hat{\boldsymbol{W}}_{j,\delta}^{N^{-1}} \boldsymbol{S})^{-1} \boldsymbol{S}' \hat{\boldsymbol{W}}_{j,\delta}^{N^{-1}}$
        Reconciled forecasts $\tilde{\boldsymbol{y}}_{i+1|\delta} = \boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{y}}_{i+1}$
        $\tilde{\boldsymbol{e}}_{i+1|\delta} = \boldsymbol{y}_{i+1} - \tilde{\boldsymbol{y}}_{i+1|\delta}$
    end for
end for
$\text{MSE}_\delta = \frac{1}{T-v}\sum_{i=v}^{T-1} (\tilde{\boldsymbol{e}}_{i+1|\delta})^2$ for each $\delta \in \Delta$
$\hat{\delta}^* = \arg\min_{\delta \in \Delta} \text{MSE}_\delta$
Compute $\hat{\lambda}^*$ on all training data using $\hat{\delta}^*$
Compute $\hat{\boldsymbol{R}}_1^*$ using $\hat{\delta}^*$ and $\hat{\lambda}^*$ on all training data, using Equation 3
Output: Estimate of optimal $\hat{\delta}^*$

3.2 PC-adjusted Estimator

To utilise the latent factors structure for better shrinkage, the PC-adjusted method takes the latent factors directly into its construction, and is appealing when there are common drivers in the time series within the hierarchy, as we saw in the Australian tourism example. It starts by decomposing the covariance matrix $\hat{\boldsymbol{W}}_1$ into a prominent principle components part (low-rank) and a orthogonal complement part $\hat{\boldsymbol{W}}_1^K$ (the correlation matrix after removing the first K principal components). Then we can apply either shrinkage or NOVELIST estimator to $\hat{\boldsymbol{W}}_1^K$:

$$\hat{\boldsymbol{W}}_1^{g,K} = \sum_{k=1}^{K} \hat{\gamma}_k \hat{\boldsymbol{\xi}}_k \hat{\boldsymbol{\xi}}_k' + g(\hat{\boldsymbol{W}}_1^K)$$

where $g(.)$ is either the shrinkage or NOVELIST estimator, $\hat{\gamma}_k$ and $\hat{\boldsymbol{\xi}}_k$ are the k-th largest eigenvalue and the corresponding eigenvector of the sample covariance matrix, respectively. Similar to the NOVELIST estimator, its PC-adjusted variant $\hat{\boldsymbol{W}}_1^{N,K}$ requires a cross-validation procedure and adjustment to obtain positive definiteness. ## Scaled Variance

To address the potential issue of the proportionality relationship $\hat{\boldsymbol{W}}_h^g = k_h g(\hat{\boldsymbol{W}}_1)$ not holding in practice, we can relax the assumption by allowing the variances to scale differently. This is done by scaling the 1-step-ahead correlation matrix back to h-step-ahead covariance matrix using the h-step-ahead standard deviations. The scaled variance estimator is given by:

$$\hat{\boldsymbol{W}}_h^{g,sv} = \boldsymbol{D}_h^{1/2} g(\hat{\boldsymbol{R}}_1) \boldsymbol{D}_h^{1/2},$$

where $\boldsymbol{D}_h = \text{diag}(\hat{\sigma}_{1,h}^2, \dots, \hat{\sigma}_{n,h}^2)$, and $\hat{\sigma}_{i,h}^2$ is the variance of the i-th series' h-step-ahead base forecast errors. Similarly, $g(.)$ is either the shrinkage or NOVELIST estimator.

If NOVELIST is used to produce h-step-ahead reconciled forecasts, the cross-validation procedure is slightly modified to evaluate the out-of-sample reconciled forecast MSE using $\hat{\boldsymbol{W}}_h^{N,sv}$ instead of $\hat{\boldsymbol{W}}_1^N$.

3.3 Constructing from h-step-ahead Residuals

Another alternative is not to rely on the assumption of proportionality at all, and directly estimate the covariance matrix from the h-step-ahead base forecast errors:

$$\hat{\boldsymbol{W}}_h^g = g(\hat{\boldsymbol{W}}_h)$$

where $\hat{\boldsymbol{W}}_h$ is the covariance matrix of in-sample h-step-ahead base forecast residuals, and $g(.)$ is either the shrinkage or NOVELIST estimator. Similar to the scaled variance approach, if NOVELIST is used, the cross-validation procedure is modified to compute $\hat{\boldsymbol{W}}_h^N$ and evaluate the out-of-sample reconciled forecast MSE using it.

Summary of MinT with Covariance Estimators

The covariance estimators explored in this paper are summarised in the table below. The abbreviations will be used in the following sections.

Covariance estimators used	Abbreviation
Shrinkage	MinT-S
NOVELIST	MinT-N
PC-adjusted Shrinkage with K PCs	MinT-S(PCK)
PC-adjusted NOVELIST with K PCs	MinT-N(PCK)
Scaled Variance Shrinkage	MinT-S(sv)
Scaled Variance NOVELIST	MinT-N(sv)
Constructed from h-step-ahead Shrinkage	MinT-S(hcov)
Constructed from h-step-ahead NOVELIST	MinT-N(hcov)

4 Evaluation of Point and Probabilistic Forecasts

This section briefly introduces the scoring rules used to evaluate the point and probabilistic forecast accuracy in Section 5 and Section 6.

For point forecasts, we use the mean squared error (MSE) to evaluate the accuracy of different reconciliation methods: $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_{i,t+h} - \tilde{y}_{i,t+h|t})^2$, where $y_{i,t+h}$ is the realised value of series i at time $t+h$, and $\tilde{y}_{i,t+h|t}$ is the reconciled point forecast.

Meanwhile, to assess the quality of the probabilistic forecasts, it is common to use proper scoring rules. A scoring rule is a function $S(.,.)$ taking a predictive distribution as its first argument and a realisation as its second argument, then returns a numerical score. We follow a convention that lower scores are better. A scoring rule is said to be *proper* if $\mathbb{E}_Q[S(Q,y)] \leq \mathbb{E}_Q[S(F,y)]$ for all F, where F is a predictive distribution produced by forecasting model, Q is the true distribution of the realisation y, and $\mathbb{E}_Q$ is the expectation with respect to Q. Hence, the expected score is minimised when the forecast distribution matches the true distribution.

We employ Winkler score and continuous ranked probability score as our univariate scoring rules, and energy score as the multivariate scoring rule. All three are proper scoring rules. In this paper, we only evaluate 1-step-ahead probabilistic forecasts, and thus we drop the subscript t and h for simplicity.

Winkler score (WS). If the $100(1-\alpha)\%$ prediction interval of i-th series is $[l_i, u_i]$ (the $\alpha/2$ and $1-\alpha/2$ quantiles), then the Winkler score is defined as:

$$WS_\alpha(l_i, u_i; y_i) = (u_i - l_i) + \frac{2}{\alpha}(l_i - y_i)\mathbf{1}(y_i < l_i) + \frac{2}{\alpha}(y_i - u_i)\mathbf{1}(y_i > u_i),$$

where y_i is the observed value of i-th series, and $\mathbf{1}(.)$ is the indicator function. The Winkler score rewards narrow intervals that contain the observation, and penalises intervals that do not contain the observation.

Continuous ranked probability score (CRPS). The CRPS is defined as the squared difference between the predictive cumulative distribution function (CDF) F_i and the empirical CDF of the observation y_i of series i:

$$CRPS(F_i, y_i) = \int_{-\infty}^{\infty} (F_i(x) - \mathbf{1}(x \geq y_i))^2 dx.$$

When the predictive distribution is Gaussian with mean μ_i and standard deviation σ_i, the CRPS has a closed-form expression:

$$CRPS(F_i, y_i) = \sigma_i \left[z_i \left(2\Phi(z_i) - 1\right) + 2\phi(z_i) - \frac{1}{\sqrt{\pi}} \right],$$

where $z_i = \frac{y_i - \mu_i}{\sigma_i}$, and $\Phi(.)$ and $\phi(.)$ are the CDF and probability density function (PDF) of a standard normal distribution, respectively.

Energy score (ES). The energy score is a multivariate generalisation of the CRPS. It is defined as:

$$ES(F, \boldsymbol{y}) = \mathbb{E}_F ||\boldsymbol{X} - \boldsymbol{y}||^\beta - \frac{1}{2} \mathbb{E}_F ||\boldsymbol{X} - \boldsymbol{X}'||^\beta,$$

where $\boldsymbol{X}$ and $\boldsymbol{X}'$ are independent random vectors with multivariate distribution F, $\boldsymbol{y}$ is the observed vector, $||.||$ is the Euclidean norm, and $\beta \in (0, 2]$. We set $\beta = 1$ following common convention.

Since the closed-form expression of the ES may not available, we approximate it using Monte Carlo samples $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ drawn from P:

$$\widehat{ES}(F, \boldsymbol{y}) = \frac{1}{M} \sum_{m=1}^{M} ||\boldsymbol{x}_m - \boldsymbol{y}|| - \frac{1}{2M(M-1)} \sum_{m=1}^{M} ||\boldsymbol{x}_m - \boldsymbol{x}_m^*||,$$

where $\boldsymbol{x}_m^*$ is a randomly selected sample from $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\} \setminus \{\boldsymbol{x}_m\}$. In our experiments, we use $M = 10000$ samples to approximate the ES.

5 Simulation

5.1 General Design

The general design of data generating process for bottom-level series is a stationary VAR(1) process, with the following structure:

$$\boldsymbol{b}_t = \boldsymbol{A}\boldsymbol{b}_{t-1} + \boldsymbol{\epsilon}_t,$$

where $\boldsymbol{A}$ is a $n_b \times n_b$ block diagonal matrix of autoregressive coefficients $\boldsymbol{A} = diag(\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m)$, with each $\boldsymbol{A}_i$ being a $n_{b,i} \times n_{b,i}$ matrix. The block diagonal structure ensures that the time series are grouped into m groups, with each group having its own autoregressive coefficients. This aim to simulate the interdependencies between the time series within each group, where reconciliation will be expected to better performed than the usual base forecasts.

The model is added with a Gaussian innovation process $\boldsymbol{\epsilon}_t$, with covariance matrix $\boldsymbol{\Sigma}$. The covariance matrix $\boldsymbol{\Sigma}$ is generated specifically using the Algorithm 1 in Hardin et al. (2013):

1. A compound symmetric correlation matrix is used for each block of size $n_{b,i}$ in $\boldsymbol{A}_i$, where the entries ρ_i for each block i are sampled from a uniform distribution between 0 and 1. They are baseline correlations within group.
2. A constant correlation, which is smaller than $\min\{\rho_1, \rho_2, \ldots, \rho_m\}$, is imposed on the entries between different blocks. It serves as baseline correlations between group.
3. The entry-wise random noise is added on top of the entire correlation matrix.
4. The covariance matrix $\boldsymbol{\Sigma}$ is then constructed by uniform sampling of standard deviations, in a range of $[\sqrt{2}, \sqrt{6}]$, for all n_b series.

We will randomly flip the signs of the covariance elements, which will create a more realistic structure in the innovation process. This can be done by pre- and post-multiplying $\boldsymbol{\Sigma}$ by a random diagonal matrix $\boldsymbol{V}$ with diagonal entries sampled from $\{-1, 1\}$, yielding $\boldsymbol{\Sigma}^* = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}$.

For all hierarchies in our experiments, we simulate two panel lengths, $T = 54$ and $T = 304$, reserving the final four observations as an out-of-sample test set. In each Monte Carlo replication ($M = 500$), we fit univariate ARIMA models (base models) to the training observations using an automatic AICc minimization algorithm from Hyndman & Khandakar (2008), implemented in the *fabletools* package (O'Hara-Wild et al., 2024), generating incoherent 1–4-step base forecasts. We then reconcile these forecasts under different methods, and evaluate their point and probabilistic forecast accuracy on the test set.

5.2 Exploring Effects of Hierarchy's Size

In our first set of experiments, we examine how MinT combined with the different estimators perform as the hierarchy expands. We generate synthetic data from the same VAR(1) framework described earlier, but vary the number of bottom-level series, n_b, across two structures: a small structure with six groups of six bottom series (n_b = 6x6 = 36), and a much larger configuration with two groups of fifty (n_b = 2x50 = 100).

In the 36-series case, each block of six forms a level-1 aggregate, and those six aggregates form the national total. The 100-series design employs a deliberately intricate aggregation path to stress-test reconciliation methods. We first sum the one hundred bottom series into ten intermediate series by grouping them in contiguous blocks of ten. These ten series are then organised into three level-2 aggregates—four, three, and four series, respectively—before finally summing to a single top node. This asymmetric hierarchy creates overlapping correlation patterns: some level-2 series share bottom-level groups, while others draw from both, emulating practical scenarios such as regional sales aggregations that span multiple product categories or overlapping territories. The aggregation paths for both structures are illustrated in Figure 5.

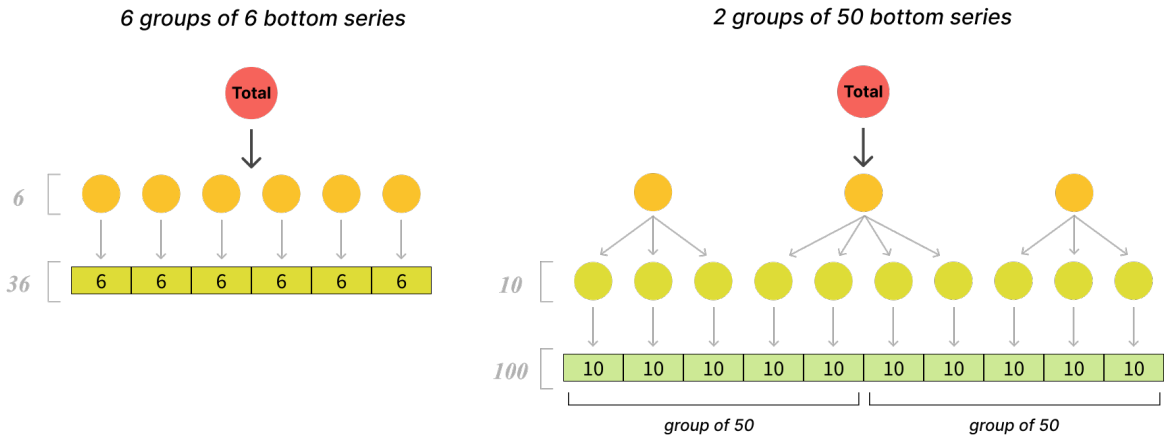

Figure 5: Aggregation structures used in the simulation experiments: 6 groups of 6 (left) and 2 groups of 50 (right)

The VAR(1) and correlation configurations for the 6 by 6 case and 2 by 50 case are illustrated in Figure 6 and Figure 7, respectively. The block diagonal structure of the VAR(1) coefficient matrices $\boldsymbol{A}$ reflects the grouping of series, and the correlation matrices- show higher correlations among series within the same group.

Figure Figure 8 illustrates the relative improvements in mean squared error (MSE) of reconciled forecasts over the incoherent base forecasts, across two structures and time series lengths. The MinT with shrinkage (*MinT-S*) and its variants are colored in mint green, while MinT with NOVELIST (*MinT-N*) are in purple. The concrete lines represent the vanilla *MinT-S*

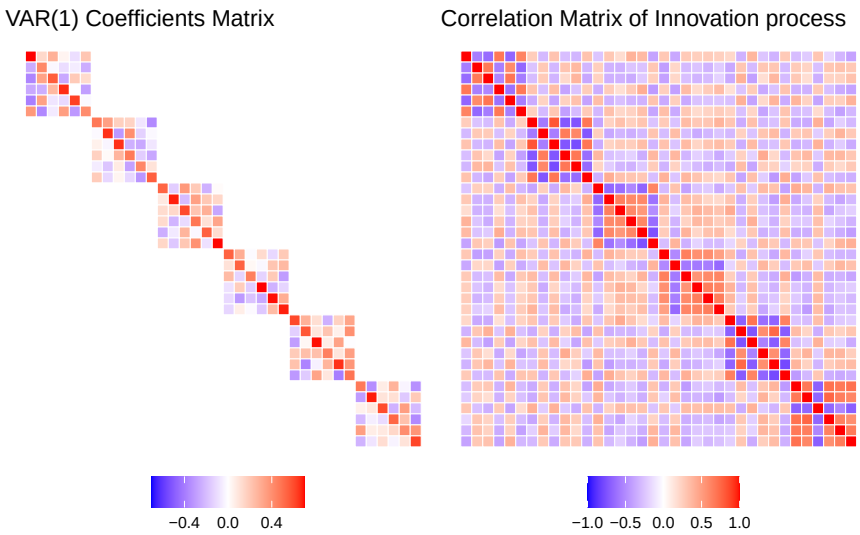

Figure 6: The VAR(1) coefficients matrix (left) and correlation matrix of the innovation process (right) for the 6 groups of 6 structure.

and *MinT-N*; the dashed lines with dot points denote the PC-adjusted variants (e.g. *MinT-S(PC1)*, *MinT-N(PC2)*); and the dotted or dashed-dotted lines indicate the scaled variance and h-step-ahead residuals versions (e.g. *MinT-S(SV)*, *MinT-N(hcov)*).

The first key observation is that methods with shrinkage slightly outperform those with NOVELIST across all scenarios, despite the differences being small. Second, the PC-adjusted variants (using one and two principal components) do not yield improvements over the vanilla versions. This is expected since the synthetic data generating process does not simulate from strong latent factors. Lastly, the scaled variance and h-step-ahead residuals approaches do not enhance performance as the forecast horizon increases, suggesting that the proportionality assumption may not be severely violated in this VAR(1) setup.

When looking into each MCMC replication, we find that there are instances where the NOVELIST estimator collapses to the shrinkage estimator due to a large optimal threshold $\hat{\delta}$ being selected in the cross-validation step, resulting in a diagonal shrinkage target.

Moving on to probabilistic forecasts, we evaluate the performance of reconciliation methods using the energy score, a proper scoring rule for multivariate predictive distributions. Figure Figure 9 presents the percentage relative improvement in energy score for 1-step-ahead forecasts. Across both hierarchies and time dimensions, the MinT methods consistently outperform the base forecasts. Meanwhile, the differences among *MinT-S* and *MinT-N* variants are insignificant, except for the 6 by 6 case with 50 observations, where shrinkage has a slight

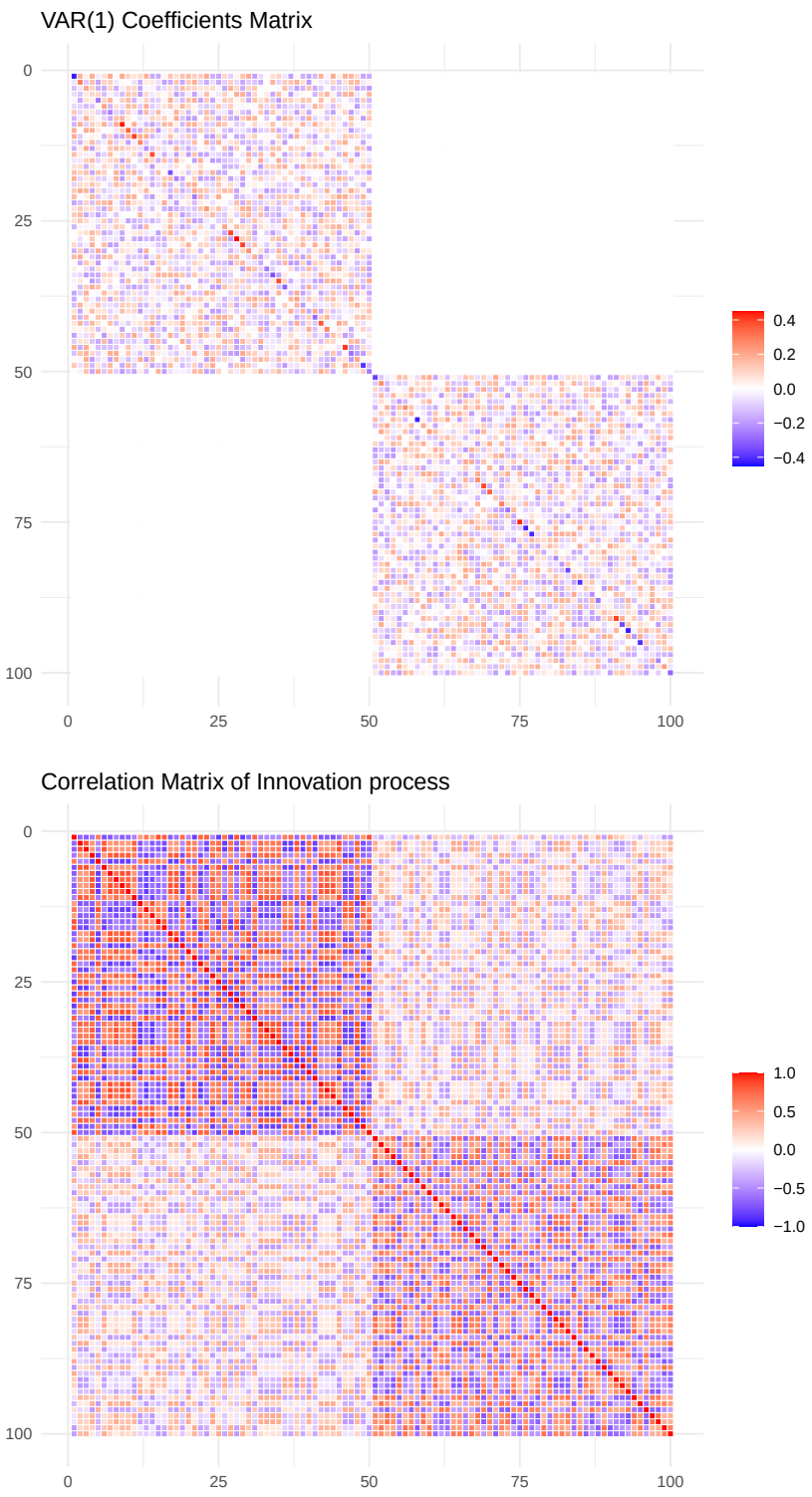

Figure 7: The VAR(1) coefficients matrix (top) and correlation matrix of the innovation process (bottom) for the 2 groups of 50 structure.

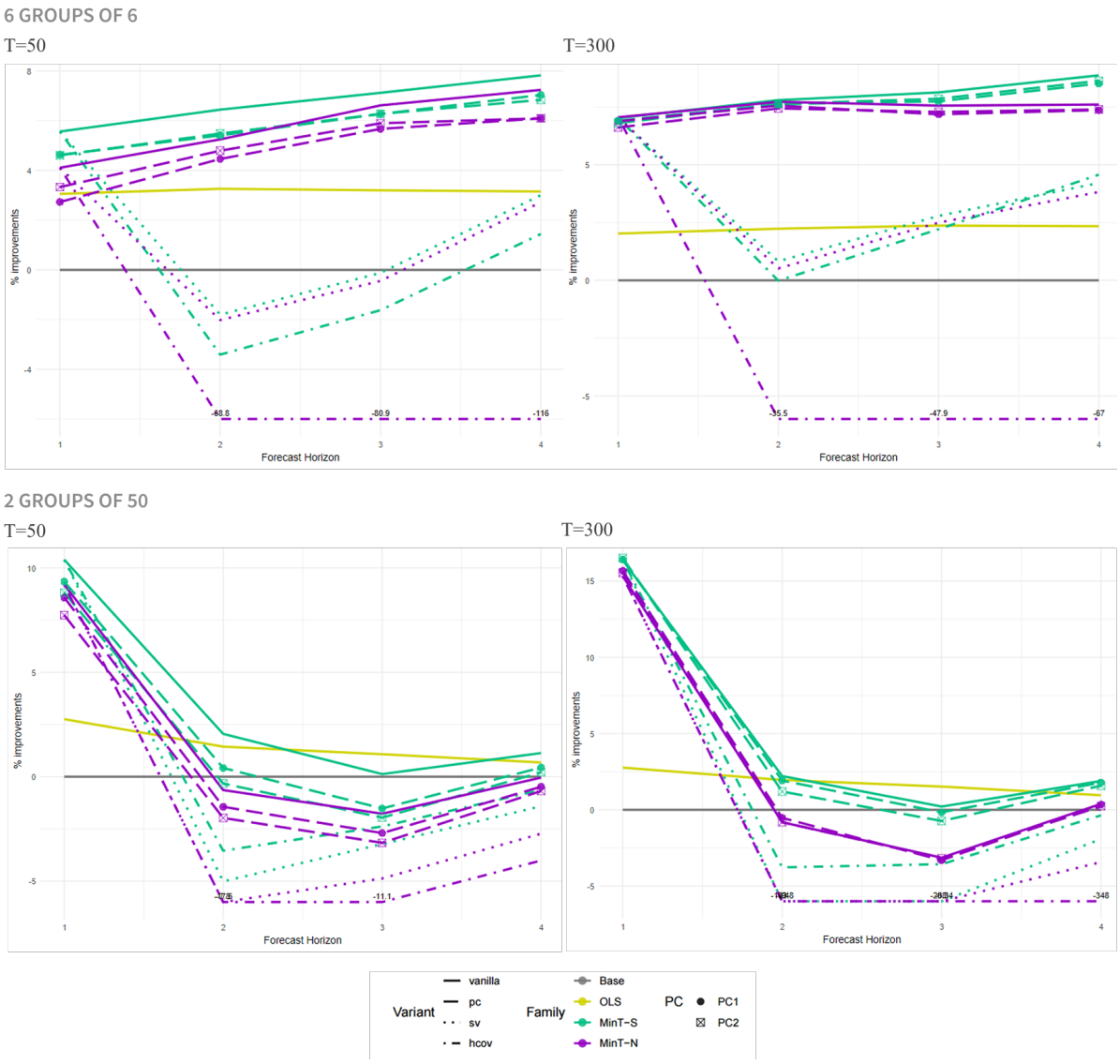

Figure 8: Percentage relative improvement in MSE of reconciled forecasts over the base forecasts in the 6 by 6 case (top row) and the 2 by 50 case (bottom row), T=50 (left column) and T=300 (right column), for 1- to 4-step-ahead forecasts. The positive (negative) entries indicate a decrease (increase) in MSE relative to base.

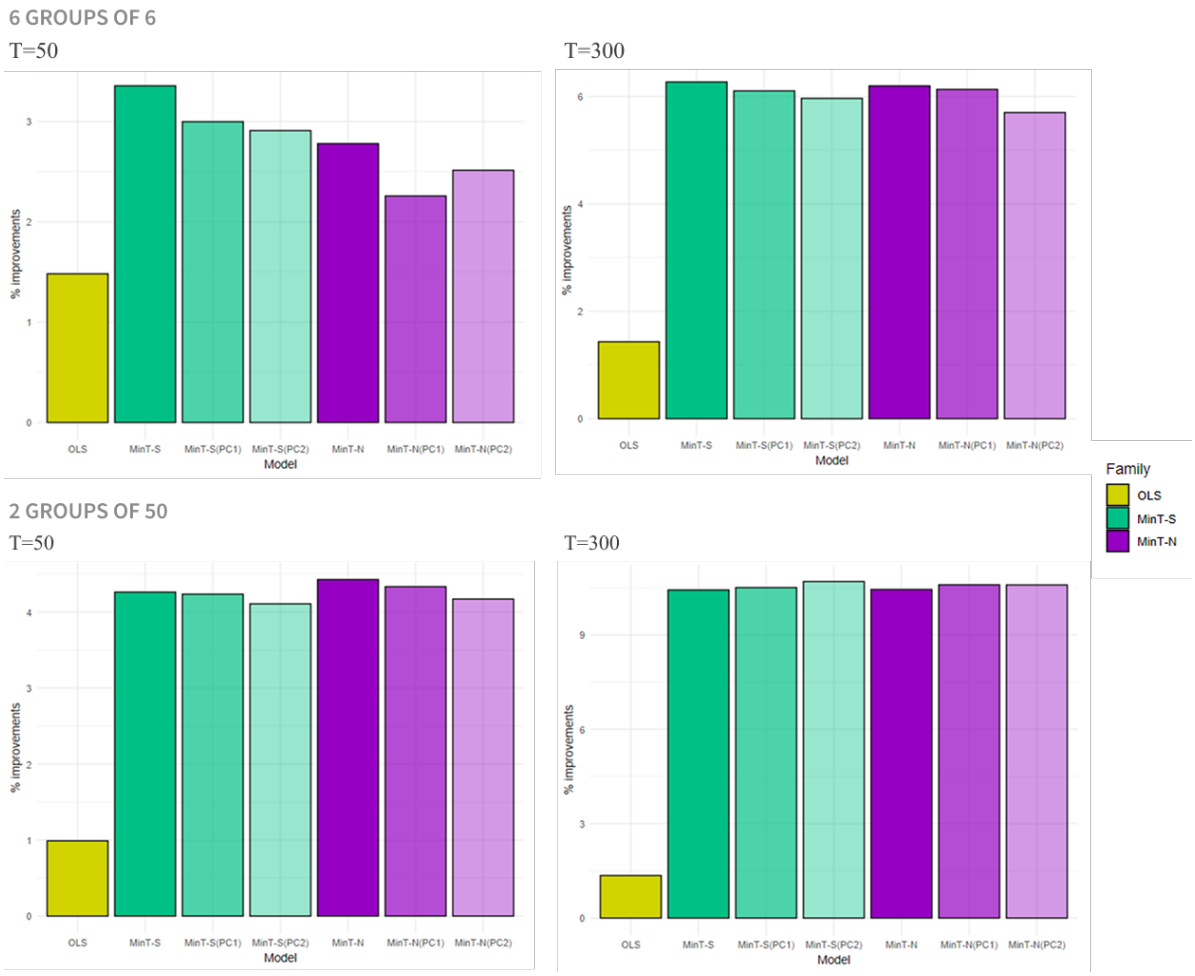

Figure 9: Percentage relative improvement in Energy score in both hierarchies and time dimensions, for 1-step-ahead forecasts. The positive entries indicate a decrease in Energy score relative to base.

edge. The PC-adjusted variants again degrade performance as we add more principal components. The scaled variance and h-step-ahead residuals approaches are not available since we only evaluate 1-step-ahead covariance estimates.

Other scoring rules, such as the Winkler score and CRPS, yield similar conclusions.

5.3 Other Data Generating Processes

In attempts to differentiate the performance of NOVELIST from the shrinkage estimator, we also simulate from a sparse covariance matrix for a 2 groups of 50 scenario, as illustrated in Figure 10. The sparse covariance matrix is obtained by randomly choosing 40% of the bottom series and setting their correlations with all other series to zero, resulting in a grid-like sparse structure. The VAR(1) coefficient matrix remains the same as in the dense case. The idea is to allow NOVELIST to exploit the sparsity in the covariance structure, since it can control the sparsity of the shrinkage target via the thresholding parameter δ. However, no profound insights can be drawn from the results.

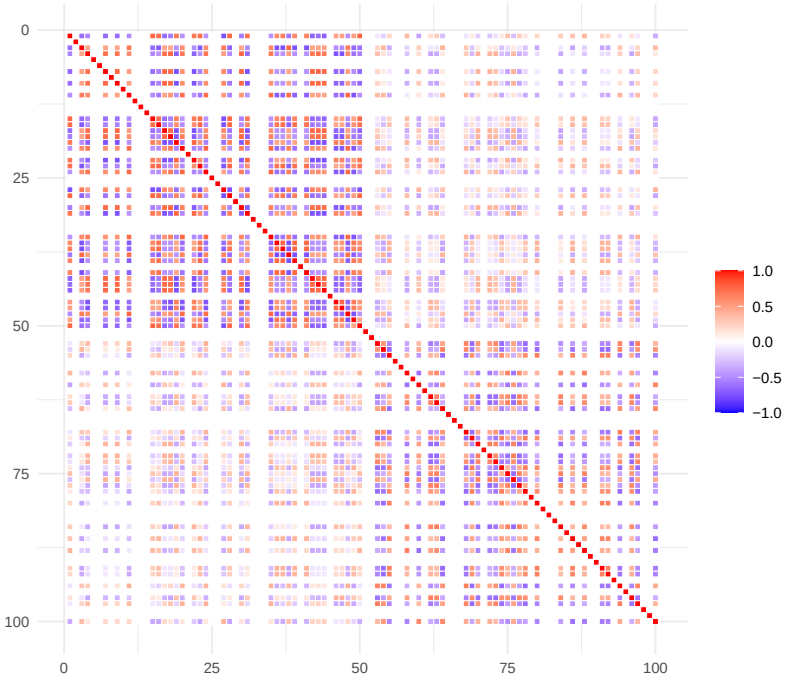

Figure 10: Sparse correlation matrix of the innovation process for 2 groups of 50 structure.

Additional designs (varying block sizes, grouped structure, aggregation paths, correlation configurations) also failed to separate NOVELIST from Shrinkage. Their nearly identical performance under these synthetic scenarios suggests that our current simulation may not unveil

the full advantages of the thresholding estimators. Nevertheless, we have not explored settings where PC variants or using h-step-ahead residuals approaches would have an edge.

These findings motivates our turn to empirical data in the next section, where latent structural features, regime shifts, and noisy, intermittent series will reveal performance differences.

6 Forecasting Australian Domestics Tourism

Domestic tourism flows in Australia exhibit a natural hierarchical and grouped structure, driven both by geography and by purpose of travel. At the top of this hierarchy lies the national total, which splits into the seven states and territories. Each state is further subdivided into tourism zones, which in turn break down into 77 regions. A complete illustration of this geographic hierarchy appears in Appendix Section 8.2. Intersecting this geographic hierarchy is a second dimension–travel motive–partitioning tourism flows into four categories: holiday, business, visiting friends and relatives, and other. Altogether, this yields a grouped system of 560 series, from the most disaggregated regional-purpose cells up to the full national aggregate. Table 2 depicts this structure.

Table 2: Hierarchical and grouped structure of Australian domestic tourism flows

Geographical division	Number of series per geographical division	Number of series per purpose	Total number of series
Australia	1	4	5
States	7	28	35
Zones	27	108	135
Regions	77	308	385
Total	112	448	560

We quantify tourism demand via "visitor nights", the total number of nights spent by Australians away from home. The data is collected via the National Visitor Survey, managed by Tourism Research Australia, using computer assisted telephone interviews from nearly 120,000 Australian residents aged 15 years and over (*Tourism Research Australia*, 2024).

The data are monthly time series spanning from January 1998 to December 2016, resulting in 228 observations per series, producing a canonical "$n \ll p$" setting which is ideal for evaluating reconciliation approaches that rely on high-dimensional covariance estimation. The extreme dimensionality over sample size mirrors many contemporary business problems, for instance, Starbucks drink sales. Tourism demand is also economically vital yet highly volatile, with geographical and purpose-specific patterns create a realistic stress-test for reconciliation algorithms.

Wickramasuriya et al. (2019) also argued that modelling spatial autocorrelations directly from the start would be challenging as in this case of a large collection of time series. Post-processing reconciliation approaches have the advantage to implicitly model this spatial autocorrelation structure, especially true for MinT.

Figure 11: Rolling-window cross-validation scheme for evaluating forecasting performance in Australia tourism data

To assess forecasting performance between models, we adopt a rolling-window cross-validation scheme. Beginning with the first 120 monthly observations (January 1998-December 2005) as the initial training set, we obtain the best-fitted ARIMA model for each of the 560 series via the automatic algorithm by minimising AICc from Hyndman & Khandakar (2008), implemented in the *fabletools* package (O'Hara-Wild et al., 2024). The 1- to 12-step-ahead base forecasts are then generated by these ARIMA models, and then reconciled using multiple approaches. To estimate the NOVELIST and its variants, we would have an extra cross-validation prodecure within this training window, as described in Section 3.1.1. We then roll the training window forward by one month and refit all models, rebuild reconciliations, and produce another batch of 1- to 12-step-ahead forecasts, repeating until the training set reaches December 2015. In total, this results in 97 out-of-sample windows. The entire procedure is illustrated in Figure 11.

Figure 12 show a main difference compared to the one of simulation results. Adjusting for a single dominant factor via PC decomposition (dashed line with dot points) tightens performance further: both *MinT-S(PC1)* and *MinT-N(PC1)* beat their unadjusted counterparts. Other than that, similar patterns are observed: among vanilla MinT variants, the shrinkage estimator (*MinT-S*) slightly outperforms NOVELIST (*MinT-N*) at most horizons; adding more than one PC brings no additional benefit, likely due to injecting estimation noise from weaker components; variants that modify the multi-step covariance, either via scaled-variance or direct h-step residual covariances, underperform standard MinT, suggesting that extra estimation at

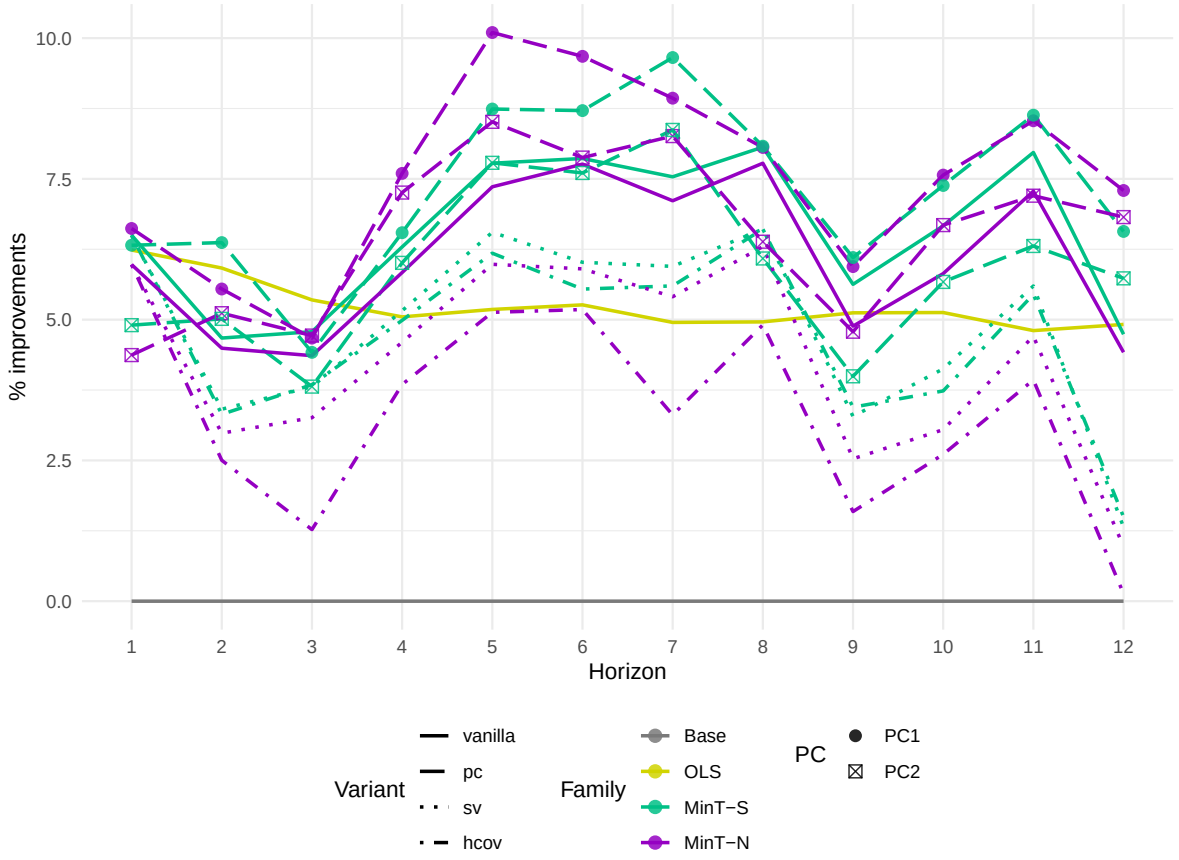

Figure 12: Percentage relative improvement in the mean squared error (MSE) of different reconciled forecasts over the base forecasts for the Australian domestic tourism data, for 1 to 12 steps ahead forecasts. The positive entries indicate an decrease in MSE.

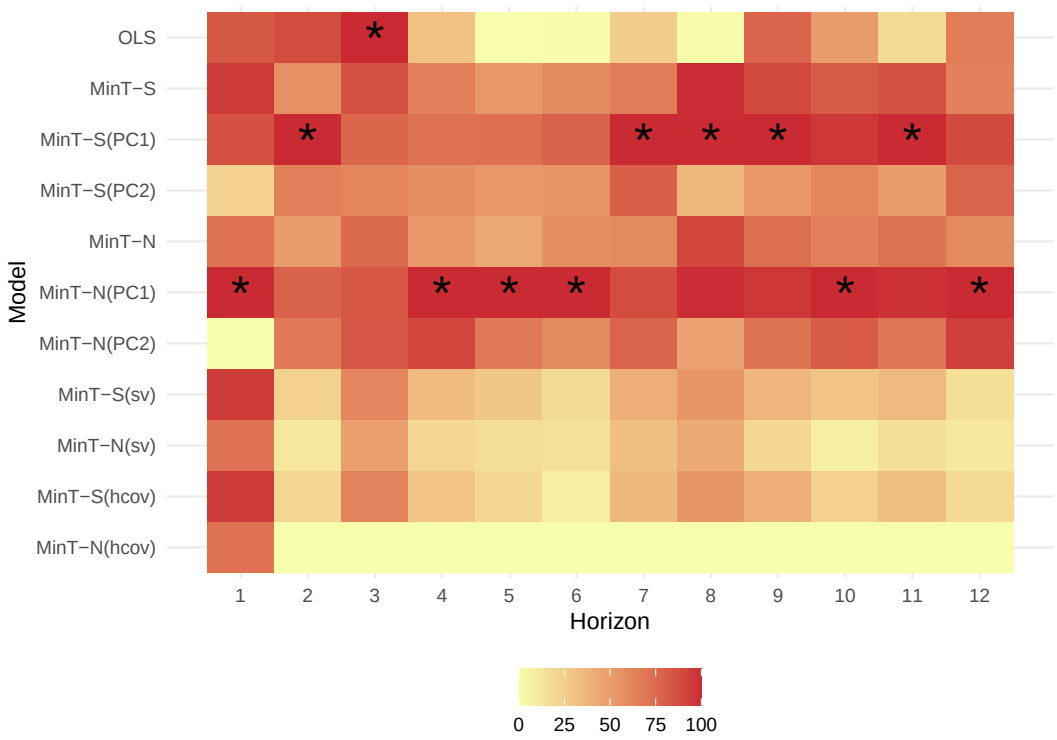

Figure 13: Heatmap of relative improvement in the mean squared error (MSE) of different reconciled forecasts over the base forecasts for the Australian domestic tourism data, for 1 to 12 steps ahead forecasts. The values are scaled to the range of 0 to 100 for better visualisation, with darker colors indicating greater improvement and best performance is noted by a star.

horizon $h > 1$ is not rewarded in this empirical analysis.

Figure 13 provides a complementary heatmap view, scaling each method's MSE improvement to a 0-100 range for better visual discrimination. Here, *MinT-N(PC1)* and *MinT-S(PC1)* emerge as the top performers across most horizons. The heatmap underscores the consistent gains from PC adjustment and highlights the diminishing returns from more complex covariance treatments.

Turning to probabilistic forecasts (1-step-ahead forecasts), Figure 14 shows that *MinT-N* (purple bars) consistently outperforms *MinT-S* (green bars) across univariate and multivariate scores. The PC1-adjusted variants again yield improvements over their vanilla counterparts, with *MinT-N(PC1)* leading overall. In the multivariate evaluation, the Energy score places *OLS* close to the PC1-adjusted MinT methods. One surprising finding is that all MinT variants underperform the base forecasts when moving to the 95% Winkler score, while *OLS* performs best.

To dissect the 95% Winkler score results, Figure 15 breaks down performance by hierarchical level, and examines empirical coverage of the 95% prediction intervals. The left panel shows that *MinT-S* underperforms the base at all levels, especially in higher aggregated levels. From our inspection, this is due to the overly shrunk variances from the shrinkage estimator, leading to narrow prediction intervals and thus high Winkler penalties when observations fall outside. The *MinT-N* method has relatively good coverage and improve the Winkler score at bottom levels, but still underperforms at higher levels. The PC-adjusted variants seem to strike a better balance, improving overall coverage and relative Winkler scores over the base.

The summary radar graph in Figure 16 consolidates these findings: among the selected MinT models by probabilistic criteria, *MinT-N(PC1)* (the yellow polygon) clearly leads across CRPS, Winkler score at 80% and 95% intervals, and Energy, extending the improvements seen in the single-metric panels. Except for the Winkler score at 95% interval, all MinT variants outperform the base forecasts.

Taken together, the evidence supports the use of reconciliation for both point and probabilistic forecasts in this high-dimensional setting. For point forecasts, MinT with shrinkage is a solid default; for probabilistic forecasts, NOVELIST performs more reliably. PC adjustment with a single dominant factor consistently enhances performance in both point and probabilistic forecast and should be considered when a dominant latent factor is evident. More complex adjustments, such as multiple PCs or horizon-specific covariances, add estimation noise without clear benefits and can be omitted for parsimony.

7 Conclusions and Future Work

This paper aims to address the limitations of the current Minimum Trace reconciliation method, a method that has garnered significant attention in both academic and practical forecasting contexts. We propose the NOVELIST estimator to address the lack of flexibility

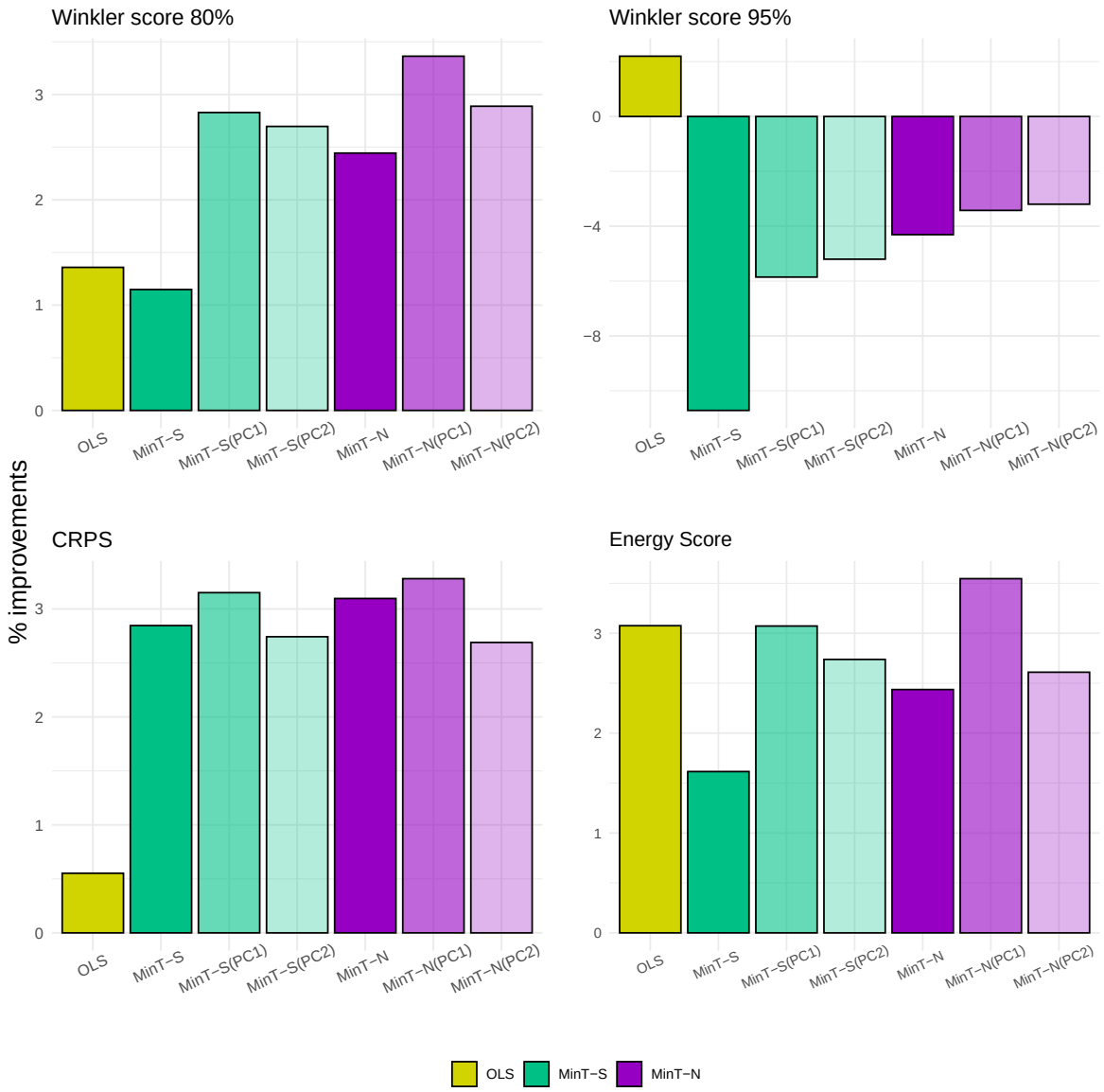

Figure 14: Percentage relative improvement in the Winkler score at 80% and 95% nominal coverage, CRPS, and Energy score of multiple reconciled forecasts over the base forecasts for the Australian domestic tourism data, for 1-step-ahead forecasts. The positive (negative) entries indicate a decrease (increase) in the probabilistic scores relative to base.

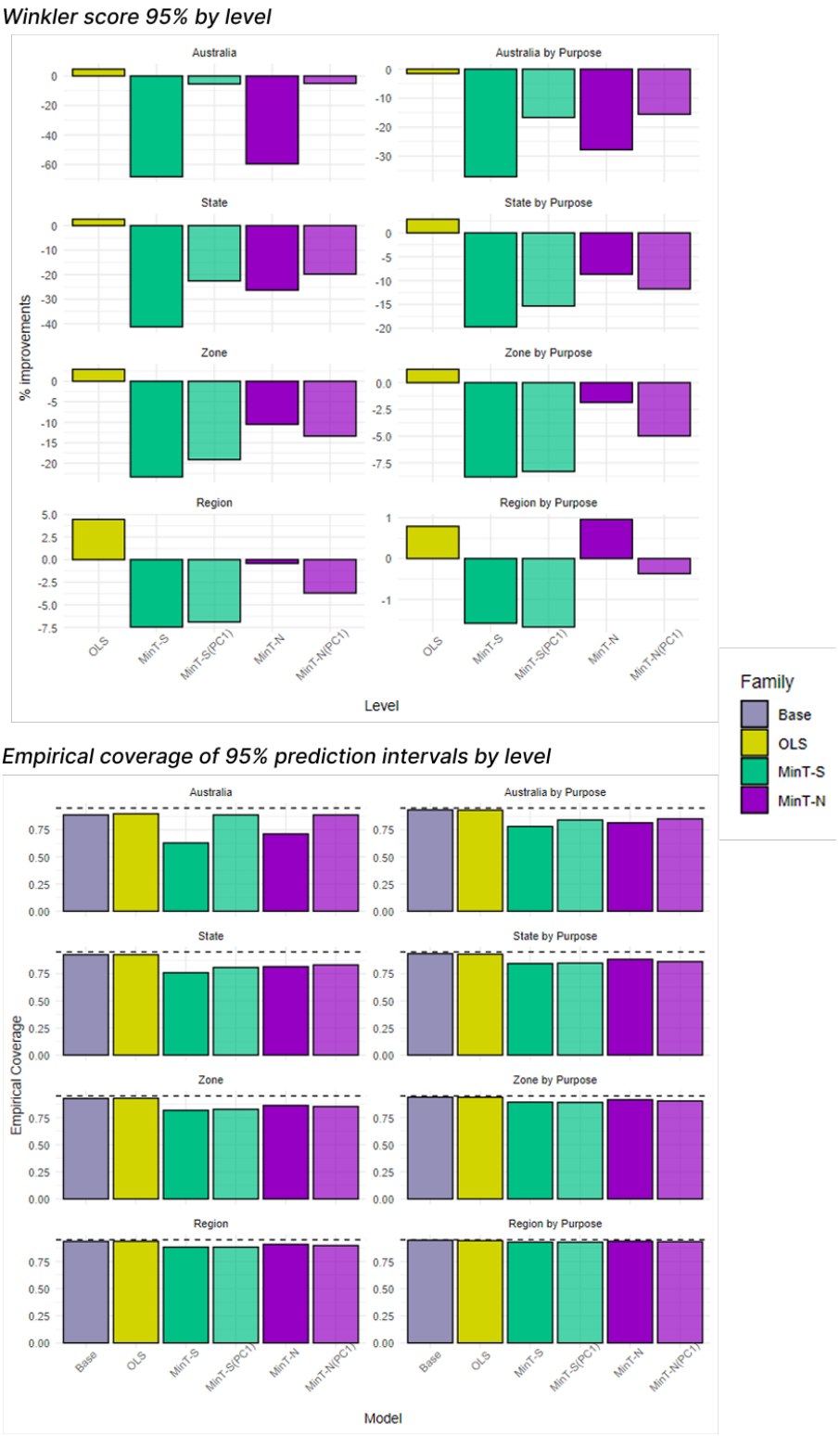

Figure 15: Percentage relative improvement in Winkler score at 95% nominal coverage by level (left), and empirical coverage of 95% prediction intervals by level (right)

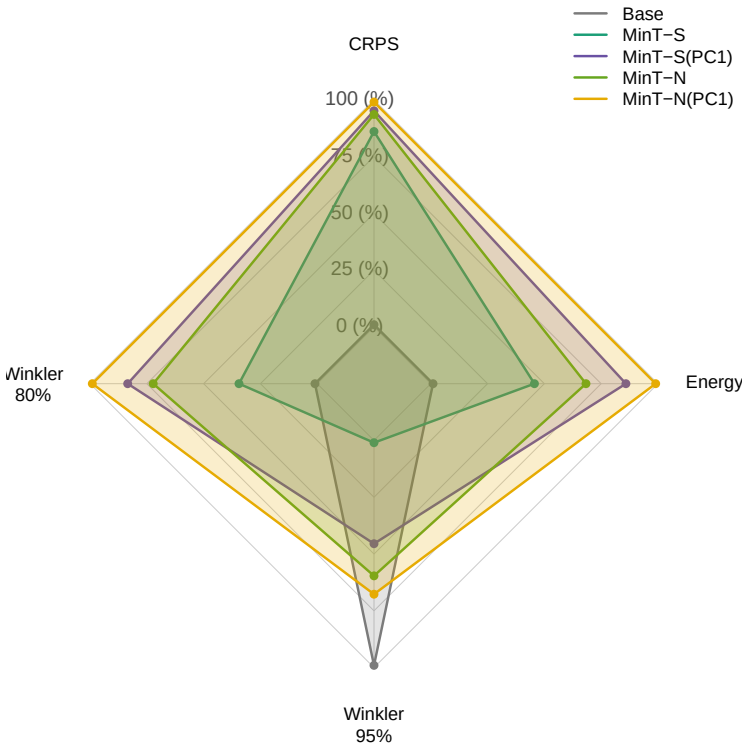

Figure 16: Radar plot of relative improvements in probabilistic scores (Winkler score at 80% and 95% intervals, CRPS, and Energy) over the base forecasts. The scores are scaled to a range of 0 to 100, with larger values indicating better performance. The outermost polygon represents the best possible score (100) and the innermost polygon represents the worst possible score (0). Only the top 4 MinT approaches are shown (with the base model).

in the uniform shrinkage target of the traditional shrinkage estimator. Additionally, we introduce PC-adjusted variants of shrinkage and NOVELIST to take into account dominant latent factors, as well as multi-step covariance variants that relax the proportionality assumption.

Empirical results on Australian domestic tourism data reveal several insights. For point forecasts, shrinkage (and its PC-adjusted version) remains a robust default choice, while NOVELIST demonstrates improved performance for probabilistic forecasts. Adjusting for a single dominant principal component consistently enhances performance across both point and probabilistic forecasts, suggesting its utility when such latent structures are present. More complex adjustments, such as incorporating multiple principal components or horizon-specific covariances, tend to introduce additional estimation noise without notable benefits.

However, several limitations and avenues for future research remain. More intricate simulation studies are necessary to further highlight the differences between shrinkage and NOVELIST, particularly in scenarios involving factor models. The current implementation does not standardise data for PC-variant methods, which may lead to higher-level series largely influencing the principal components. Future work could also explore alternative methods that leverage cross-series information or eigenvalue/eigenvector structures. Additionally, understanding why MinT with shrinkage underperforms in the Winkler 95% score, especially at higher aggregation levels, needs further investigation. Finally, evaluating h-step-ahead probabilistic forecasts remains an open area for further research.

All data generation, covariance estimation, and reconciliation routines were implemented in the ReconCov R package and is available under an open-source license on GitHub (Su, 2025).

8 Appendix

8.1 Appendix: Simulation Supplementary

8.2 Appendix: Australian Domestic Tourism Geographical Hierarchy

Table 3: Geographical divisions of Australia.

Series	Name	Label	Series	Name	Label
1	Australia	Total	57	Bundaberg	CAA
2	NSW	A	58	Capricorn	CAB
3	NT	B	59	Fraser Coast	CAC
4	QLD	C	60	Gladstone	CAD
5	SA	D	61	Mackay	CAE
6	TAS	E	62	Southern Queensland Country	CAF
7	VIC	F	63	Outback Queensland	CBA
8	WA	G	64	Brisbane	CCA
9	ACT	AA	65	Gold Coast	CCB
10	Metro NSW	AB	66	Sunshine Coast	CCC
11	Nth Coast NSW	AC	67	Townsville	CDA
12	Nth NSW	AD	68	Tropical North Queensland	CDB
13	Sth Coast NSW	AE	69	Whitsundays	CDC
14	Sth NSW	AF	70	Clare Valley	DAA
15	Central NT	BA	71	Flinders Ranges and Outback	DAB
16	Nth Coast NT	BB	72	Murray River, Lakes and Coorong	DAC
17	Central Coast QLD	CA	73	Riverland	DAD
18	Inland QLD	CB	74	Adelaide	DBA
19	Metro QLD	CC	75	Adelaide Hills	DBB
20	Nth Coast QLD	CD	76	Barossa	DBC
21	Inland SA	DA	77	Fleurieu Peninsula	DCA
22	Metro SA	DB	78	Kangaroo Island	DCB
23	Sth Coast SA	DC	79	Limestone Coast	DCC
24	West Coast SA	DD	80	Eyre Peninsula	DDA
25	Nth East TAS	EA	81	Yorke Peninsula	DDB
26	Nth West TAS	EB	82	East Coast	EAA
27	Sth TAS	EC	83	Launceston and the North	EAB
28	East Coast VIC	FA	84	North West	EBA
29	Metro VIC	FB	85	West Coast	EBB
30	Nth East VIC	FC	86	Hobart and the South	ECA
31	Nth West VIC	FD	87	Gippsland	FAA
32	West Coast VIC	FE	88	Lakes	FAB
33	Nth WA	GA	89	Phillip Island	FAC
34	Sth WA	GB	90	Geelong and the Bellarine	FBA
35	West Coast WA	GC	91	Melbourne	FBB
36	Canberra	AAA	92	Peninsula	FBC
37	Central Coast	ABA	93	Central Murray	FCA
38	Sydney	ABB	94	Goulburn	FCB
39	Hunter	ACA	95	High Country	FCC
40	North Coast NSW	ACB	96	Melbourne East	FCD

41	Blue Mountains	ADA	97	Murray East	FCE
42	Central NSW	ADB	98	Upper Yarra	FCF
43	New England North West	ADC	99	Ballarat	FDA
44	Outback NSW	ADD	100	Bendigo Loddon	FDB
45	South Coast	AEA	101	Central Highlands	FDC
46	Capital Country	AFA	102	Macedon	FDD
47	Riverina	AFB	103	Mallee	FDE
48	Snowy Mountains	AFC	104	Spa Country	FDF
49	The Murray	AFD	105	Western Grampians	FDG
50	Alice Springs	BAA	106	Wimmera	FDH
51	Barkly	BAB	107	Great Ocean Road	FEA
52	Lasseter	BAC	108	Australia's North West	GAA
53	MacDonnell	BAD	109	Australia's Golden Outback	GBA
54	Darwin	BBA	110	Australia's Coral Coast	GCA
55	Katherine Daly	BBB	111	Australia's South West	GCB
56	Litchfield Kakadu Arnhem	BBC	112	Destination Perth	GCC

References

Angam, B., Beretta, A., De Poorter, E., Duvinage, M., & Peralta, D. (2025). Forecast reconciliation for vaccine supply chain optimization. In *Communications in computer and information science* (pp. 101–118). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-74650-5/_6

Athanasopoulos, G., Ahmed, R. A., & Hyndman, R. J. (2009). Hierarchical forecasts for australian domestic tourism. *International Journal of Forecasting*, *25*(1), 146–166. https://doi.org/10.1016/j.ijforecast.2008.07.004

Athanasopoulos, G., Hyndman, R. J., Kourentzes, N., & Panagiotelis, A. (2024). *Forecast reconciliation: A review*. *40*(2), 430–456. https://www.sciencedirect.com/science/article/pii/S0169207023001097

Ben Taieb, S., & Koo, B. (2019). Regularized regression for hierarchical forecasting without unbiasedness conditions. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. https://doi.org/10.1145/3292500.3330976

Cai, T., & Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, *106*(494), 672–684. https://doi.org/10.1198/jasa.2011.tm10560

Carrara, C., Zambon, L., Azzimonti, D., & Corani, G. (2025). A novel shrinkage estimator of the covariance matrix for hierarchical time series. In *Italian statistical society series on advances in statistics* (pp. 140–145). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-96736-8/_24

Di Modica, C., Pinson, P., & Ben Taieb, S. (2021). Online forecast reconciliation in wind power prediction. *Electric Power Systems Research*, *190*(106637), 106637. https://doi.org/10.1016/j.epsr.2020.106637

El Gemayel, J., Lafarguette, R., Itd, K. M., et al. (2022). *United arab emirates: Technical assistance reportliquidity management and forecasting*.

Erven, T. van, & Cugliari, J. (2015). Game-theoretically optimal reconciliation of contemporaneous hierarchical time series forecasts. In *Modeling and stochastic learning for forecasting in high dimensions* (pp. 297–317). Springer International Publishing. https://doi.org/10.1007/978-3-319-18732-7/_15

Fan, J., Liao, Y., & Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *75*(4), 603–680. https://doi.org/10.1111/rssb.12016

Gamakumara, P. (2020). *Probabilistic forecast reconciliation: Theory and applications* [PhD thesis, Monash University]. https://doi.org/10.26180/5e4ca9d0c4b9d

Hardin, J., Garcia, S. R., & Golan, D. (2013). A method for generating realistic correlation matrices. *The Annals of Applied Statistics*, *7*(3), 1733–1762. https://www.jstor.org/stable/23566492

Higham, N. (2002). Computing the nearest correlation matrix—a problem from finance. *Ima Journal of Numerical Analysis*, *22*, 329–343. https://doi.org/10.1093/IMANUM/22.3.329

Huang, N., & Fryzlewicz, P. (2019). NOVELIST estimator of large correlation and covariance

matrices and their inverses. *Test (Madrid, Spain)*, *28*(3), 694–727. https://doi.org/10.1007/s11749-018-0592-4

Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., & Shang, H. L. (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis*, *55*(9), 2579–2589. https://doi.org/10.1016/j.csda.2011.03.006

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, *27*, 1–22. https://doi.org/10.18637/JSS.V027.I03

Hyndman, R. J., Lee, A. J., & Wang, E. (2016). Fast computation of reconciled forecasts for hierarchical and grouped time series. *Computational Statistics & Data Analysis*, *97*, 16–32. https://doi.org/10.1016/j.csda.2015.11.007

Ledoit, O., & Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, *10*(5), 603–621. https://doi.org/10.1016/s0927-5398(03)00007-0

Ledoit, O., & Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Annals of Statistics*. https://doi.org/10.1214/12-AOS989

Ledoit, O., & Wolf, M. (2020). Analytical nonlinear shrinkage of large-dimensional covariance matrices. *Annals of Statistics*, *48*(5), 3043–3065. https://doi.org/10.1214/19-AOS1921

Li, H., Li, H., Lu, Y., & Panagiotelis, A. (2019). A forecast reconciliation approach to cause-of-death mortality modeling. *Insurance, Mathematics & Economics*, *86*, 122–133. https://doi.org/10.1016/j.insmatheco.2019.02.011

Nixtla. (2025). *Time series forecasting software*. https://www.nixtla.io/

O'Hara-Wild, M., Hyndman, R. J., & Wang, E. (2024). *Fabletools R package* (Version v0.5.0). https://fabletools.tidyverts.org/

Panagiotelis, A., Gamakumara, P., Athanasopoulos, G., & Hyndman, R. J. (2023). Probabilistic forecast reconciliation: Properties, evaluation and score optimisation. *European Journal of Operational Research*, *306*(2), 693–706. https://doi.org/10.1016/j.ejor.2022.07.040

Schäfer, J., & Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, *4*(1), Article32. https://doi.org/10.2202/1544-6115.1175

Seaman, B., & Bowman, J. (2022). Applicability of the M5 to forecasting at walmart. *International Journal of Forecasting*, *38*(4), 1468–1472. https://doi.org/10.1016/j.ijforecast.2021.06.002

Su, V. (2025). *ReconCov R package* (Version beta). https://github.com/lordtahdus/ReconCov

Tourism research australia. (2024). https://www.tra.gov.au/.

Wickramasuriya, S. L. (2024). Probabilistic forecast reconciliation under the gaussian framework. *Journal of Business & Economic Statistics: A Publication of the American Statistical Association*, *42*(1), 272–285. https://doi.org/10.1080/07350015.2023.2181176

Wickramasuriya, S. L., Athanasopoulos, G., & Hyndman, R. J. (2019). Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *Journal of the American Statistical Association*, *114*(526), 804–819. https://doi.org/10.1080/01621459.2018.1448825

Wickramasuriya, S. L., Turlach, B. A., & Hyndman, R. J. (2020). Optimal non-negative

forecast reconciliation. *Statistics and Computing*, *30*(5), 1167–1182. https://doi.org/10.1007/s11222-020-09930-0