Enhancing Forecast Reconciliation: A Study of Alternative Covariance Estimators

Vincent Su

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), POET (Principal Orthogonal complEment Thresholding), and others.

1 Introduction

In time series forecasting, aggregation occurs in a variety of settings. While a formal definition of hierarchical time series can be found in Section 2.1, we can think of Starbucks sales data as an illustrative 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, followed by national sales for each country, then city sales for each city within a country, and finally outlet sales for each outlet in a city. As a result, there are over 40,000 individual outlet sales to forecast, plus additional series at higher levels of aggregation such as city and country. 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.

This hierarchical structure is not unique to the Starbucks sales data; it can be found in many other domains, such as national tourism, electricity demand, or Gross Domestic Product

(GDP). The impact of methods for forecasting hierarchical time series has not been limited to academia, with industry also showing a strong interest. Many companies and organisations have adopted these methods in practice, including Amazon, the International Monetary Fund, IBM, SAP, and more (Athanasopoulos et al., 2024).

- Talk about the history and evolution of forecasting hierarchical time series, starting from the early heuristic methods to the modern statistical approaches.
    - Single level methods
    - Optimal combination methods (OLS, WLS)
    - MinT
    - Bayesian, Machine learning
    - Probabilistic methods

Traditionally, forecasting these hierarchical time series has been done using single-level methods, such as bottom-up, top-down, and middle-out approaches. Bottom-up methods involve generating forecasts for the bottom-level series and aggregating them to higher levels. Top-down methods start with forecasts for the only top-level series and disaggregate them down. Middle-out methods combine both approaches by forecasting middle-level series and then aggregating or disaggregating as needed. Despite their simplicity, these methods only anchor forecasts to a single level, implying a large loss of information on the hierarchy's inherent correlation structure. Additionally, the most disaggregated series often are very noisy or even intermittent, and the higher-level data might be smoother due to the aggregation. Furthermore, as we saw from the Starbucks example considering the sales of different kinds of drinks at each aggregation level – formally defined as grouped structure in Section 2.1 – the disaggregation becomes more complex since the disaggregation paths are not unique. Consequently, these single-level methods tend to give poor results across other levels of the hierarchy.

To overcome these issues, forecast reconciliation was introduced by Hyndman et al. (2011), and later developed by Hyndman et al. (2016), Wickramasuriya et al. (2019), and ... to achieve coherency in point forecasts and enhance accuracy. Forecast reconciliation projects a collection of independent base forecasts into a set of coherent forecasts that respect the linear constraints defining a hierarchical or grouped time-series system.

- Discuss the interests in MinT and how it has become a standard method for forecast reconciliation.
- Discuss the MinT's reliance on a good estimate of the covariance matrix of base forecast errors and other gaps
    - Empirical evidence of MinT under perform
    - Comparison with other methods
- Explain why this paper focus on exploring alternative covariance estimators for MinT. And is there any paper talk about this.

  - Is better estimate of W_h really lead to better performance?

- Talk about the paper outline and what will be covered in the following sections.

2 Theoretical Framework

2.1 Hierarchical tructure

The hierarchical structure can be represented as a tree, as shown in Figure 1. The top level of the tree represents the total value of all series, while the lower levels represent the series at different levels of disaggregation. When there are attributes of interest that are crossed, such as the Starbucks drinks sales at any aggregation level (brand-wise, national, city, or outlet) is also considered by kinds of drinks (e.g., coffees, refreshers), the structure is described as a grouped time series. As illustrated in Figure 2, the aggregation or disaggregation paths are not unique.

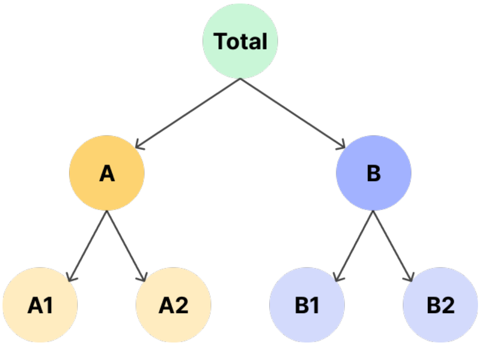

Figure 1: A 2-level hierarchical tree structure

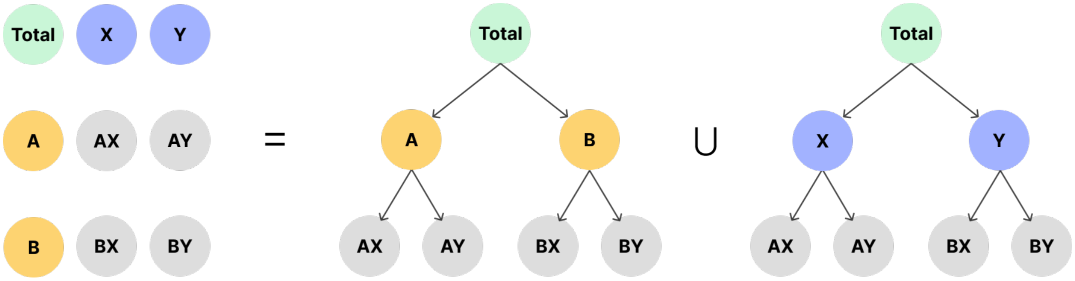

Figure 2: A 2-level grouped structure, which can be considered as the union of two hierarchical trees with common top and bottom level series

For simplicity, we refer to both of these structures as hierarchical time series, we will distinguish between them if and when it is necessary. All hierarchical structures can be represented using matrix algebra:

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

where $\boldsymbol{S}$ is a summing matrix of order $n \times n_b$ which aggregates the bottom-level series $\boldsymbol{b}_t \in \mathbb{R}^{n_b}$ to the series at aggregation levels above. The vector $\boldsymbol{y_t} \in \mathbb{R}^n$ contains all observations at time t. The summing matrix $\boldsymbol{S}$ for the tree structure in Figure 1 is:

$$\boldsymbol{S} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ & \multicolumn{2}{c}{\boldsymbol{I_4}} & \end{bmatrix}.$$

Assume we produce h-step-ahead base forecasts $\hat{\boldsymbol{b}}_{t+h|t}$ for the bottom-level series, obtained by any prediction methods. Then pre-multiplying them by $\boldsymbol{S}$ we get:

$$\tilde{\boldsymbol{y}}_{t+h|t} = \boldsymbol{S}\hat{\boldsymbol{b}}_{t+h|t}\,. \tag{1}$$

We refer to $\tilde{\boldsymbol{y}}_{t+h|t}$ as coherent forecasts, as they respect the aggregation structure. We also refer to this way of obtaining coherent forecasts by summing the bottom-level forecasts as the bottom-up approach. However, generating forecasts this way is anchored only to prediction models at a single level, and will not be utilising the inherent information from other levels. This drawback applies to the top-down and middle-out approaches. For example, the bottom-level data can be very noisy or even intermittent, and the higher-level data might be smoother due to the aggregation.

Another issue with expressing reconciled methods as in Equation 1 is that it restricts the reconciliation to only single-level approaches. Thus, Hyndman et al. (2011) suggested a generalised expression for all existing methods, which also provides a framework for new methods to be developed:

$$\tilde{\boldsymbol{y}}_{t+h|t} = \boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{y}}_{t+h|t}\ , \tag{2}$$

for a suitable $n_b \times n$ matrix $\boldsymbol{G}$. $\boldsymbol{G}$ maps 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 $\boldsymbol{S}$. The choice of $\boldsymbol{G}$ determines the composition of reconciled forecasts $\tilde{\boldsymbol{y}}_{t+h|t}$, and modern reconciliation methods are developed to estimate $\boldsymbol{G}$.

2.2 The Minimum Trace (MinT) Reconciliation

Wickramasuriya et al. (2019) framed the problem as minimising the variances of all reconciled forecast errors $\text{Var}[y_{t+h} - \tilde{y}_{t+h|t}] = \boldsymbol{S}\boldsymbol{G}\boldsymbol{W}_h\boldsymbol{G}'\boldsymbol{S}'$, where $\boldsymbol{W}_h = \mathbb{E}(\hat{\boldsymbol{e}}_{t+h|t}\ \hat{\boldsymbol{e}}'_{t+h|t})$ is the positive definite covariance matrix of the h-step-ahead base forecast errors. They showed that this is equivalent to minimising the trace of the reconciled forecast error covariance matrix (sum of the diagonal elements - the variances). The Minimum Trace (MinT) solution is given by

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

Wickramasuriya et al. (2019) also showed that MinT is an algebraic generalisation of the GLS, and the OLS and WLS methods are special cases of MinT when $\boldsymbol{W}_h$ is an identity matrix I_{n_b} and a diagonal matrix $\text{diag}(\boldsymbol{W}_h)$, respectively. In this paper, we place our main focus on the MinT method.

The MinT solution hinges on a reliable, positive-definite estimate of $\boldsymbol{W}_h$, which is challenging to estimate in high-dimensional setting. The sample covariance matrix is unstable and non-positive-definite when the number of series n is huge and larger than the time dimension T. To tackle this issue, the original paper Wickramasuriya et al. (2019) adopted the diagonal-target shrinkage estimator from Schäfer & Strimmer (2005), given by

$$\hat{\boldsymbol{W}}_1^{shr} = \lambda_D\hat{\boldsymbol{W}}_{1,D} + (1-\lambda_D)\hat{\boldsymbol{W}}_1\,,$$

where $\hat{\boldsymbol{W}}_{1,D}$ is a diagonal matrix comprising the diagonal entries $\text{diag}(\hat{\boldsymbol{W}}_1)$. We refer to any $\lambda \in [0,1]$ as the shrinkage intensity parameter, the subscript specifies which 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_D:

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

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.

However, 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 error covariance matrix $\hat{\boldsymbol{W}}_1$. The twenty largest eigenvalues of the covariance matrix are plotted in Figure 3. We can see that the point of inflexion occurs at the component with 5th largest eigenvalue, indicating a prominent principal components structure.

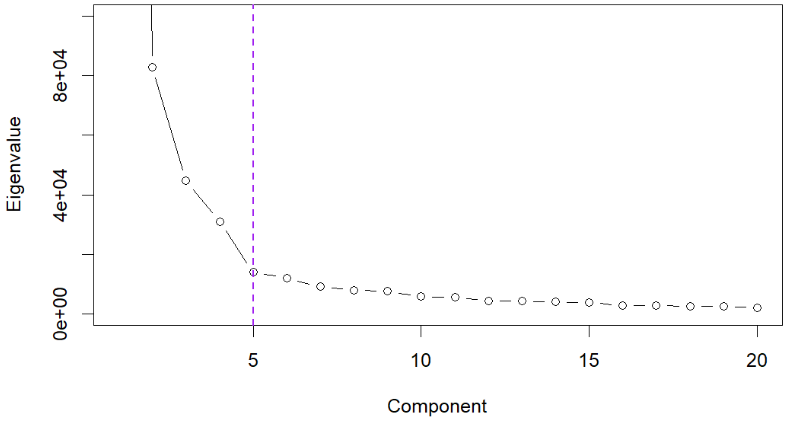

Figure 3: Twenty largest eigenvalues of one-step-ahead in-sample base forecast error covariance, Australian domestic overnight trips

Additionally, the shrinkage estimator shrinks all off-diagonal elements towards zeros with equal weights λ_D. We might prefer to better preserve strong signals, and largely reduce the effects of small, noisy correlations. In the next sections, we will explore several options that take these two issues into account.

3 Covariance Estimation Approaches

3.1 NOVELIST Estimator

The NOVELIST (NOVEL Integration of the Sample and Thresholded Covariance) estimator, proposed by Huang & Fryzlewicz (2019), is currently the main focus of this research project. It 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. However, it does not guarantee to be positive definite.

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:

$$\hat{\boldsymbol{W}}_1^N = \lambda_\delta \hat{\boldsymbol{W}}_{1,\delta} + (1-\lambda_\delta)\hat{\boldsymbol{W}}_1, \tag{3}$$

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:

$$\hat{\boldsymbol{R}}_1^N = \lambda_\delta \hat{\boldsymbol{R}}_{1,\delta} + (1-\lambda_\delta)\hat{\boldsymbol{R}}_1, \tag{4}$$

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

$$\hat{r}_{1,ij}^{\delta} = \operatorname{sign}(\hat{r}_{1,ij}) \max(|\hat{r}_{1,ij}| - \delta,\ 0), \tag{5}$$

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:

$$\hat{\lambda}(\delta) = \frac{\sum_{i\neq j} \widehat{Var}(\hat{r}_{1,ij})\ \mathbf{1}(|\hat{r}_{1,ij}| \leq \delta)}{\sum_{i\neq j}(\hat{r}_{1,ij} - \hat{r}_{1,ij}^{\delta})^2}, \tag{6}$$

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

On the other hand, the optimal threshold δ^* 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 $\hat{\delta^*}$, with the corresponding $\hat{\lambda}^*$ and $\hat{\boldsymbol{R}}_1^N(\hat{\delta}^*, \hat{\lambda}^*)$, that minimises the average out-of-sample reconciled forecast errors. The formal algorithm is given in the Section 7.2 Appendix. Although it is not required to fit forecasting models multiple times, the cross-validation procedure is still computationally expensive as it computes the NOVELIST estimator and perform reconciliation for each threshold value.

Note that when $\delta \in \left[\max_{i\neq j}|\hat{r}_{1,ij}|,\ 1\right]$, the NOVELIST estimator collapses to the shrinkage estimator, and when $\delta = 0$, it becomes the sample covariance matrix.

3.2 POET

The POET (Principal Orthogonal complEment Thresholding) estimator, proposed by Fan et al. (2013), is another "sparse" + "non-sparse" covariance estimator. It 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.

The POET method starts by decomposing the correlation matrix $\hat{\boldsymbol{R}}_1$ into a prominent principle components part (low-rank) and a orthogonal complement part $\hat{\boldsymbol{R}}_{1,K}$ (the correlation matrix after removing the first K principal components). Then it applies thresholding to $\hat{\boldsymbol{R}}_{1,K}$. The POET estimator is given by:

$$\hat{\boldsymbol{R}}_1^K = \sum_{k=1}^{K} \hat{\gamma}_k \hat{\boldsymbol{\xi}}_k \hat{\boldsymbol{\xi}}_k' + T(\hat{\boldsymbol{R}}_{1,K})$$

where $\hat{\gamma}_k$ and $\hat{\boldsymbol{\xi}}_k$ are the kth largest eigenvalue and the corresponding eigenvector of the sample covariance matrix, respectively, and $T(.)$ is the thresholding function, which can be either soft-thresholding, hard-thresholding, or others.

3.3 PC-adjusted NOVELIST

This approach is best of both worlds, leveraging the strengths of both NOVELIST and POET. The PC-adjusted (Principal-Component-adjusted) NOVELIST overcomes the shortcomings of the current shrinkage estimator, taking prominent PCs into account while also offers extra flexibility. The idea is to apply the NOVELIST estimator to the orthogonal complement part $\hat{\boldsymbol{R}}_{1,K}$, and then add the principal components part back. The PC-adjusted NOVELIST estimator is formulated as:

$$\hat{\boldsymbol{R}}_1^{N,K} = \sum_{k=1}^{K} \hat{\gamma}_k \hat{\boldsymbol{\xi}}_k \hat{\boldsymbol{\xi}}_k' + \hat{\boldsymbol{R}}_{1,K}^N \ ,$$

where $\hat{\boldsymbol{R}}_{1,K}^N$ is the NOVELIST estimator applied to the orthogonal complement part $\hat{\boldsymbol{R}}_{1,K}$. Similar to the NOVELIST estimator, $\hat{\boldsymbol{R}}_1^{N,K}$ is not guaranteed to be positive definite.

Methods to ensure positive definiteness of the NOVELIST estimator (and its PC-adjusted variant) will be explored and studied in the project. Huang & Fryzlewicz (2019) proposed to diagonalise the NOVELIST estimator and replace any eigenvalues that fall under a certain small positive threshold by the value of that threshold. Alternatively, we can implement the algorithm of Higham (2002) that computes the nearest positive definite matrix to a given matrix.

4 Simulation

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 $\boldsymbol{\Sigma}^* = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}$.

4.1 Exploring Effects of Hierarchy's Size

Using the data generating process described above, we consider three different 2-level hierarchical structures, with the bottom-level series n_b being 2 groups of 2 (4 series), 6 groups of 6 (36 series), and 2 groups of 50 (100 series), respectively. The first structure is exactly the same as the one in Figure 1, with 2 series at the level 1. The second structure has 36 bottom series and are aggregated in groups of size 6 to form 6 level-1 series, which are finally aggregated to form the top level.

The third structure is larger, we construct two groups of fifty series each for the bottom level. However, rather than collapsing each group into a single aggregate as in previous hierarchies, we impose a more intricate, multi-level aggregation path to stress-test reconciliation methods. We first form ten intermediate series by summing bottom-level series in consecutive blocks of ten. These ten series are then assigned to three second-level aggregates–groups of four, three, and four–and finally summed to produce a single top-level series. This asymmetric hierarchy creates overlapping correlation patterns: some level-2 series share bottom-level groups, while others draw from both.

The VAR(1) coefficients matrices are block diagonal and would have entries between groups being zeros. To save space, we will only illustrate the settings of the second structure (6 groups of 6) in Figure 4, and the illustrations of the first and third structure is attached in Section 7.1 Appendix.

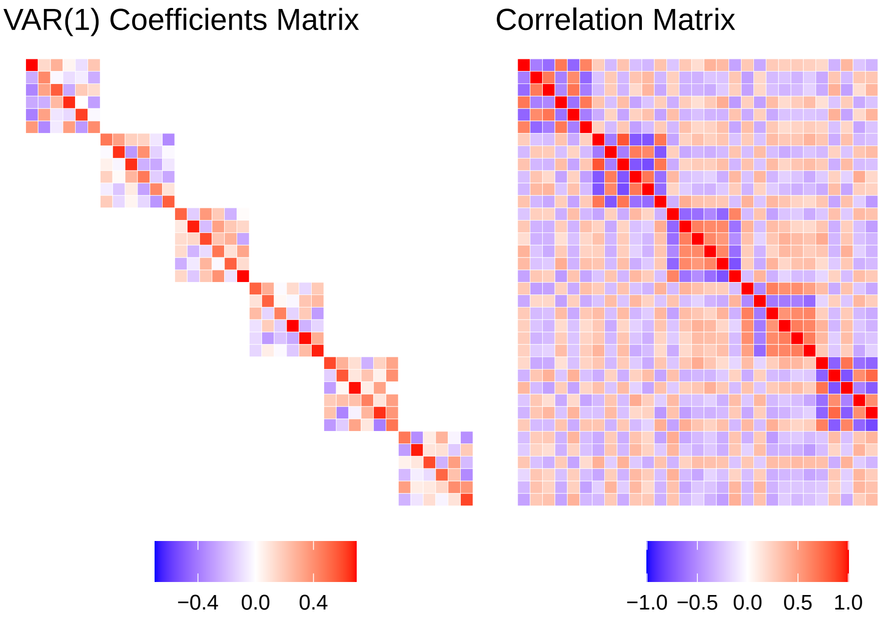

Figure 4: Heatmaps of the VAR(1) coefficient matrix and correlation matrix for the 50x2 structure.

For each series, $T = 54$ and 304 observations are generated. The first 50 and 300 observations are used for training, and the last 4 observations are used for testing. The training data is used to compute the best fitted ARIMA models by minimising the AICc criterion, in which we use the automatic algorithm from Fabletools R package (O'Hara-Wild et al., 2024). We refer to them as base models, and their base forecasts are then reconciled using the MinT with different covariance estimators. These include using the unbiased sample covariance matrix - mint_sample, the shrinkage estimator - mint_shr, and the NOVELIST estimator - mint_n. The Monte Carlo simulation is repeated $M = 200$ times, in which the parameters for data generating process is fixed.

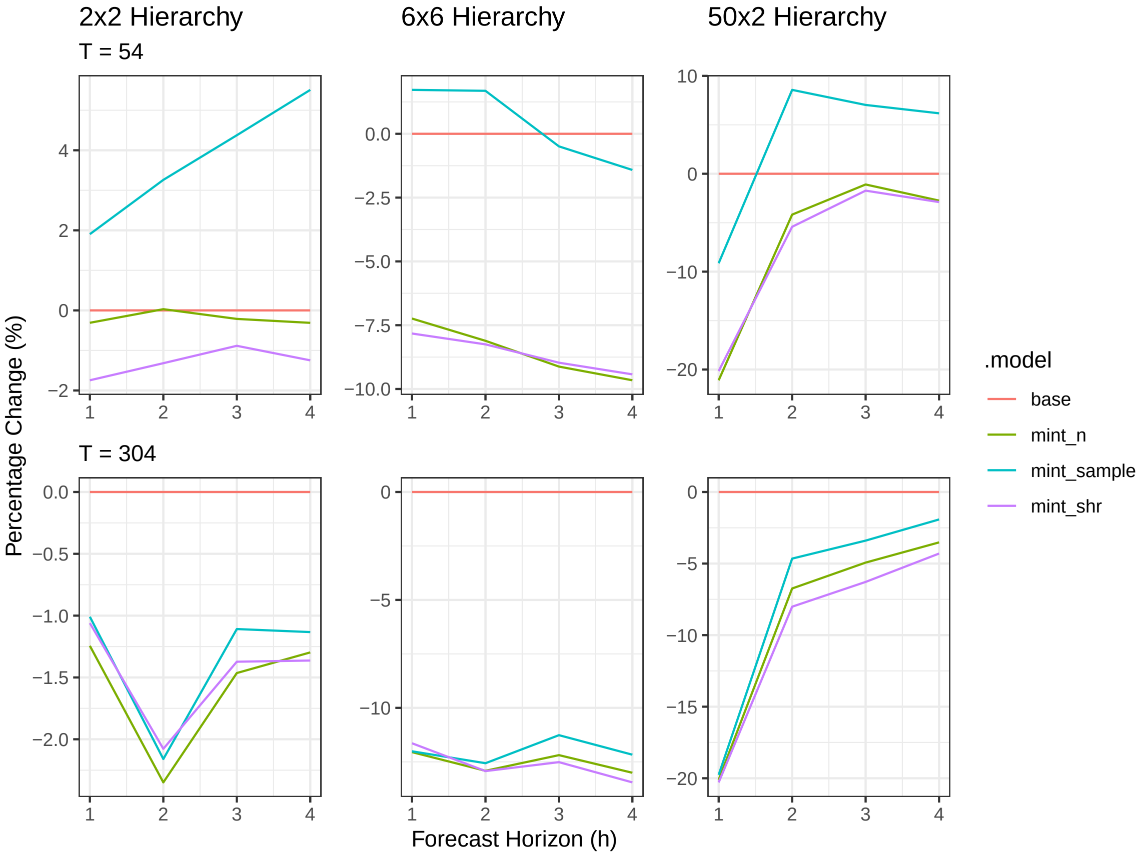

Figure 5: Relative improvement of the MSE of reconciled forecasts over the base forecasts for the 2x2, 6x6, and 50x2 hierarchical structures, for 1 to 4 steps ahead forecasts, with 2 time series lengths (T = 54 and T = 304).

In this setting, we aim to assess the performance of MinT with NOVELIST when the size of the hierarchy becomes larger. As shown in Figure 5, ...

4.2 Exploring the sparsity of the DGP covariance matrix

In our second simulation study, we design a data-generating process that contrasts "dense" and "sparse" correlation regimes among bottom-level series, reflecting settings one might encounter in practice. Specifically, we construct two groups of fifty series each, with strong within-group dependencies throughout and either modest between-group correlations (the dense scenario) or complete independence (the sparse scenario). These correlation matrices are depicted in Figure Figure 6. Both scenarios share the same VAR(1) coefficient structure as in our previous simulations; only the innovation covariance changes. Such a setup mirrors real-world contexts where, for example, sales within a product line may exhibit strong co-movements, while those in a separate line operate nearly independently.

Rather than collapsing each group into a single aggregate as in previous practice, we impose a more intricate, multi-level aggregation path to stress-test reconciliation methods. We first form ten intermediate series by summing bottom-level series in consecutive blocks of ten. These ten series are then assigned to three second-level aggregates–groups of four, three, and four–and finally summed to produce a single top-level series. This asymmetric hierarchy creates overlapping correlation patterns: some level-2 series share bottom-level groups, while others draw from both.

Figure Figure 7 presents out-of-sample mean squared error improvements over the base forecasts for each reconciliation strategy under both dense and sparse settings, with two panel lengths–54 and 304 observations–reserving the last four points for testing. In both short and long samples, MinT using either the shrinkage estimator (mint_shr) or the NOVELIST estimator (mint_n) delivers pronounced gains over incoherent ARIMA forecasts, particularly at the one-step horizon where cross-series correlations most directly inform the forecast adjustments. Although the two MinT variants perform almost indistinguishably overall, mint_n edges out mint_shr in the immediate horizon, whereas mint_shr slightly outperforms for longer horizons. By contrast, MinT with the raw sample covariance (mint_sample) suffers in small-sample settings; as expected, its performance improves dramatically with 304 data points, since the sample covariance becomes more reliable with larger n. This highlights the practical necessity of regularized estimators in high-dimensional, low-sample contexts, a situation common in real applications where histories are short relative to the number of series.

We also experimented with alternative DGPs: varying group sizes, aggregation paths, and VAR(1) structure. However, to our surprise, most designs failed to distinguish mint_shr from mint_n in any substantive way. Their nearly identical performance under these synthetic scenarios suggests that our current simulation may not unveil the full advantages of the thresholding estimators. This finding motivates our turn to empirical data, where latent

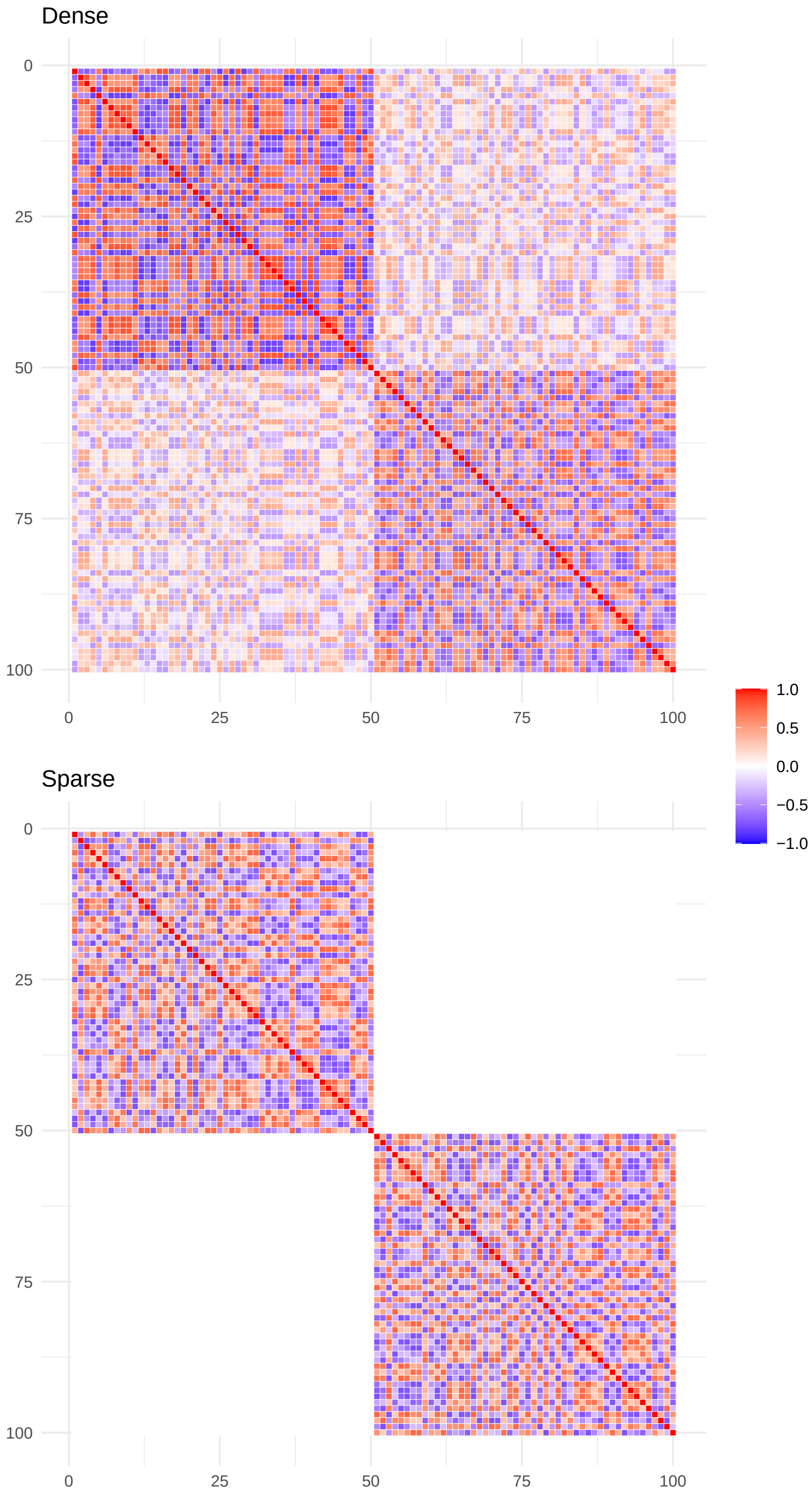

Figure 6: Heatmaps of the dense and sparse correlation matrix of the data generating process.

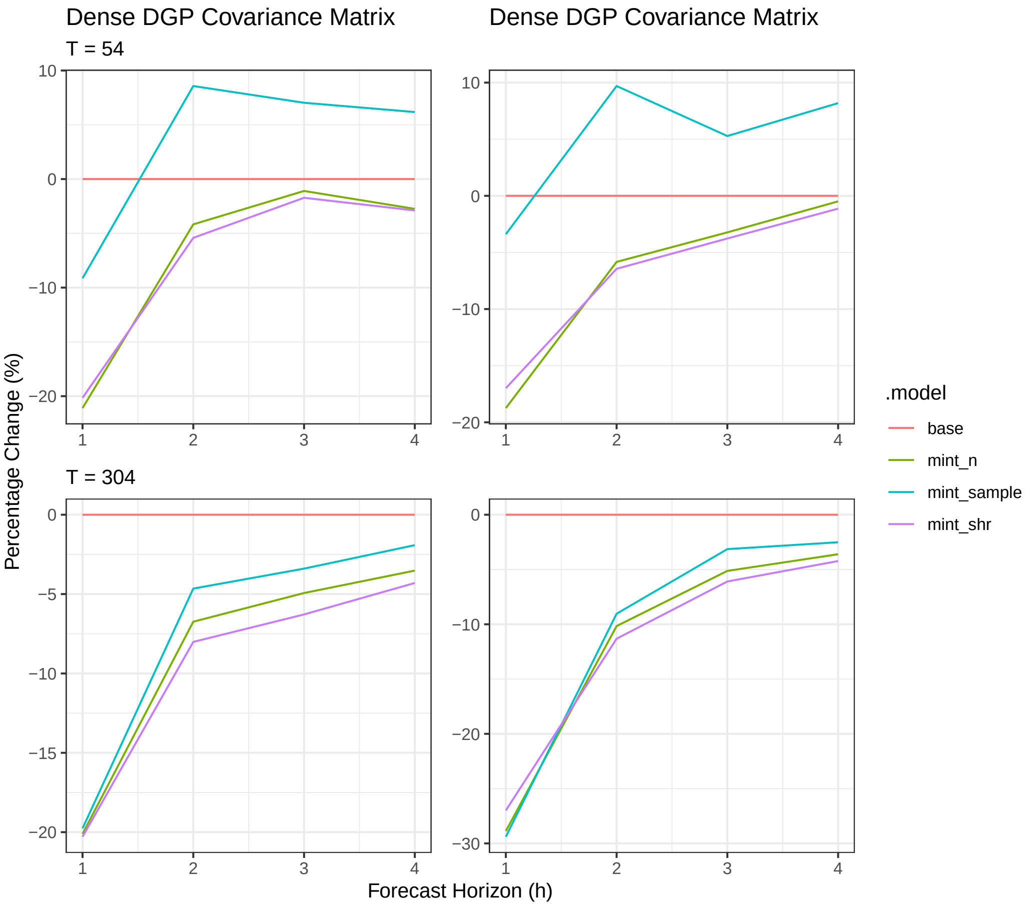

Figure 7: Relative improvement of the MSE of reconciled forecasts over the base forecasts for the 50x2 hierarchical structure with dense and sparse DGP's correlation matrix, for 1 to 4 steps ahead forecasts, with 2 time series lengths (T = 54 and T = 304).

structural features and regime shifts, which we will discuss in the next section, may reveal performance differences.

5 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 sub-divided into tourism zones, which in turn break down into 77 regions. 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.

- A table/figure to illustrate the hierarchy

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 2019, resulting in 264 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. Consequently, this panel offers both a rich policy case study and a stringent statistical laboratory for comparing reconciliation strategies that exploit cross-series information to improve forecasts when historical data are scarce.

Wickramasuriya et al. (2019) also argued that modelling spatial autocorrelations would be challenging as in the case of a large collection of time series. Reconciliation approaches have the advantage to implicitly model this spatial autocorreltion structure, especially when the MinT method is used.

To assess forecasting performance, we adopt a rolling-window cross-validation scheme. Beginning with the first 96 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 in the fabletools package, by minimising AICc (O'Hara-Wild et al., 2024). The 1- to 12-step-ahead base forecasts are then generated by these ARIMA models. These incoherent base forecasts are reconciled via OLS, MinT with shrinkage (mint_shr), and MinT with the NOVELIST estimator (mint_n). We then roll the training window forward by one month

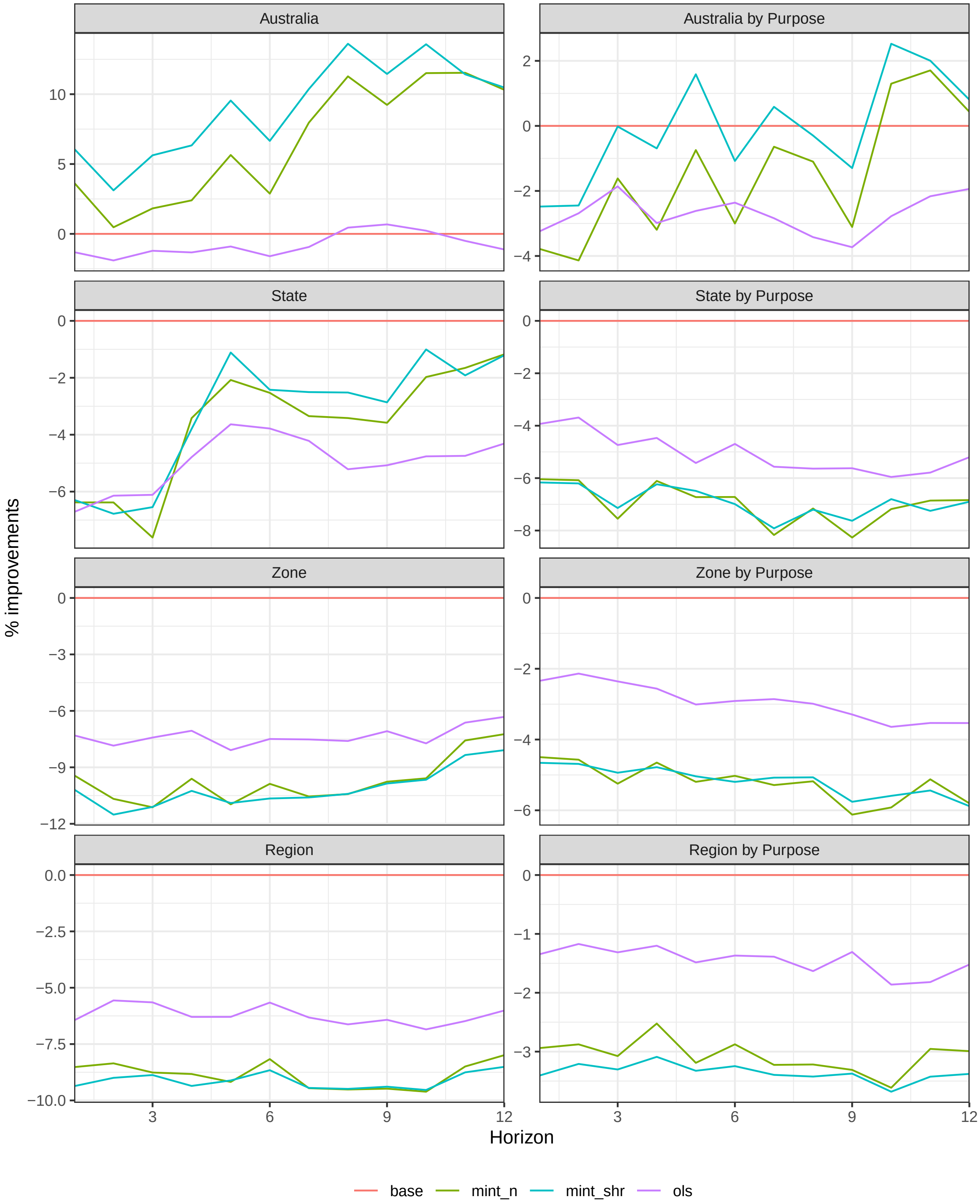

Figure 8: Relative improvement of the MSE of reconciled forecasts over the base forecasts for the Australian domestic tourism flows

and refit all models, rebuild reconciliations, and produce another batch of 1- to 12-step-ahead forecasts, repeating until December 2018. In total, this results in 156 out-of-sample windows and an equal number of forecast sets for each series.

Note that the number of series is larger than the number of observations (560 compared to 96), hence the sample covariance matrix is not positive definite and will not be considered.

The results are presented in Figure 8, which shows the relative improvement of the MSE of reconciled forecasts over the base forecasts for the Australian domestic tourism flows. The results are grouped by levels of aggregation. MinT variants show improvement over the base ARIMA forecasts for middle to lower aggregation levels. We hardly differentiate the point accuracy performance between mint_shr and mint_n in these level, except at the most disaggregated, where mint_shr slightly edges out mint_n.

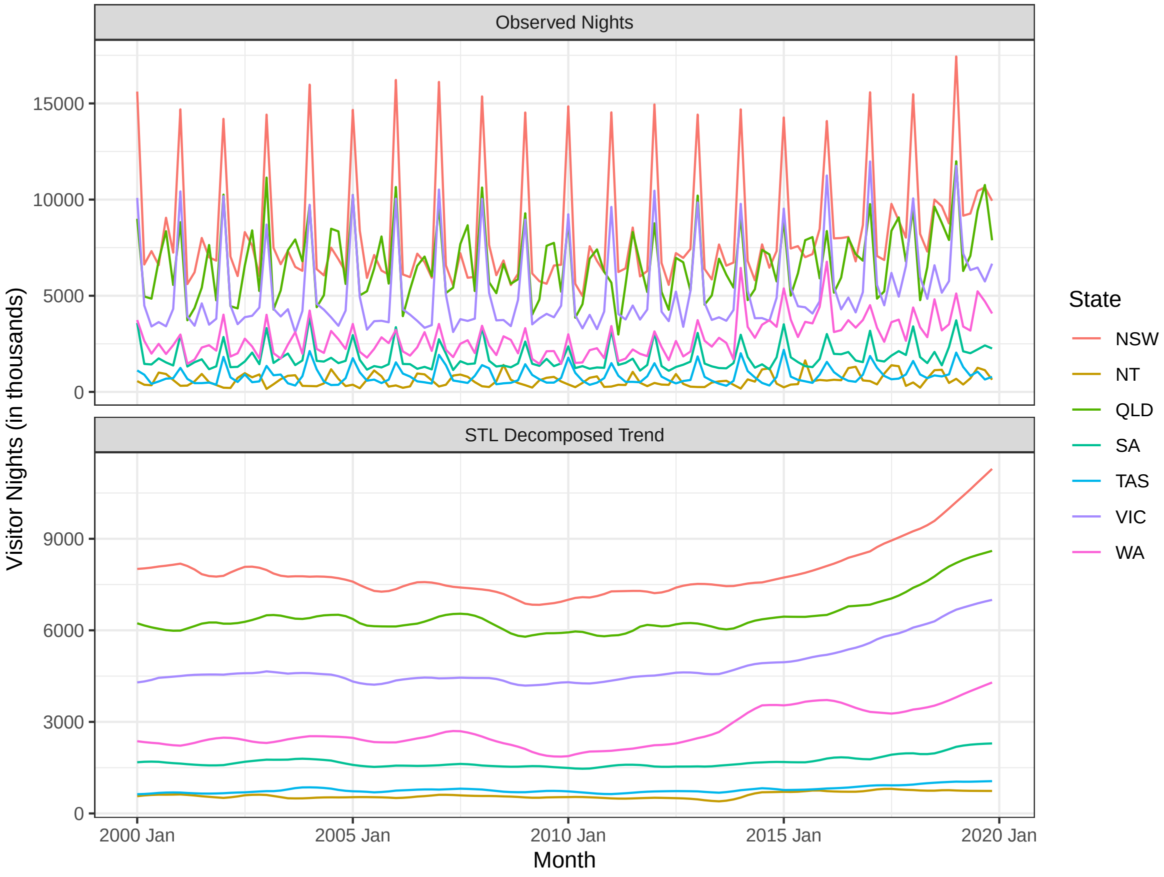

Figure 9: Monthly domestic overnight tourism nights in Australia by state, with STL decomposition to extract trend component

Surprisingly, however, MinT underperforms OLS at higher aggregation level and even the base forecasts at country-wise. This contradicts with the canonical results from Wickramasuriya et al. (2019), prompting further investigation. Applying STL decomposition by Bandara et al. (2022) to state-aggregated observations reveals a structural change around 2016: four large states (New South Wales, Victoria, Queensland, Western Australia) develop a modest upward trend in visitor nights, as shown in Figure 9. The trend is likely due to the increasing popularity of tourism in these places, as well as the impact of various factors such as economic growth, population growth, and changes in consumer preferences.

Neither ARIMA nor reconciliation strategies capture this shift, degrading accuracy at the most aggregated level. From the results in Figure 10, excluding the post-2016 period (i.e., restricting the final training window to end on December 2015) restores reconciliation's performance: both MinT methods now clearly outperform OLS and the base forecasts at all levels. Additionally, we notice that mint_n increasingly outperforms mint_shr at higher levels (Australia, State, and Australia by Purposes levels). This is an advantage of NOVELIST estimator. At higher levels most series are combinations of many bottom cells, their cross-series base forecast error correlations are strong and genuinely useful. NOVELIST keeps these strong links at those higher levels, while its thresholding feature minimises the effects of weak, noisy ones.

The empirical analysis on real, hierarchical Australian domestic tourism shreds lights on the strengths of the NOVELIST estimator–advantages that were not fully apparent in previous simulation studies. Going forward, we plan to embed this structure into our data-generating processes and to pinpoint contexts where NOVELIST's thresholding yields gains beyond those of Shrinkage. Moreover, our findings underscore the need to accommodate structural changes–the post-2016 Australian domestic tourism demand shifts we encountered.

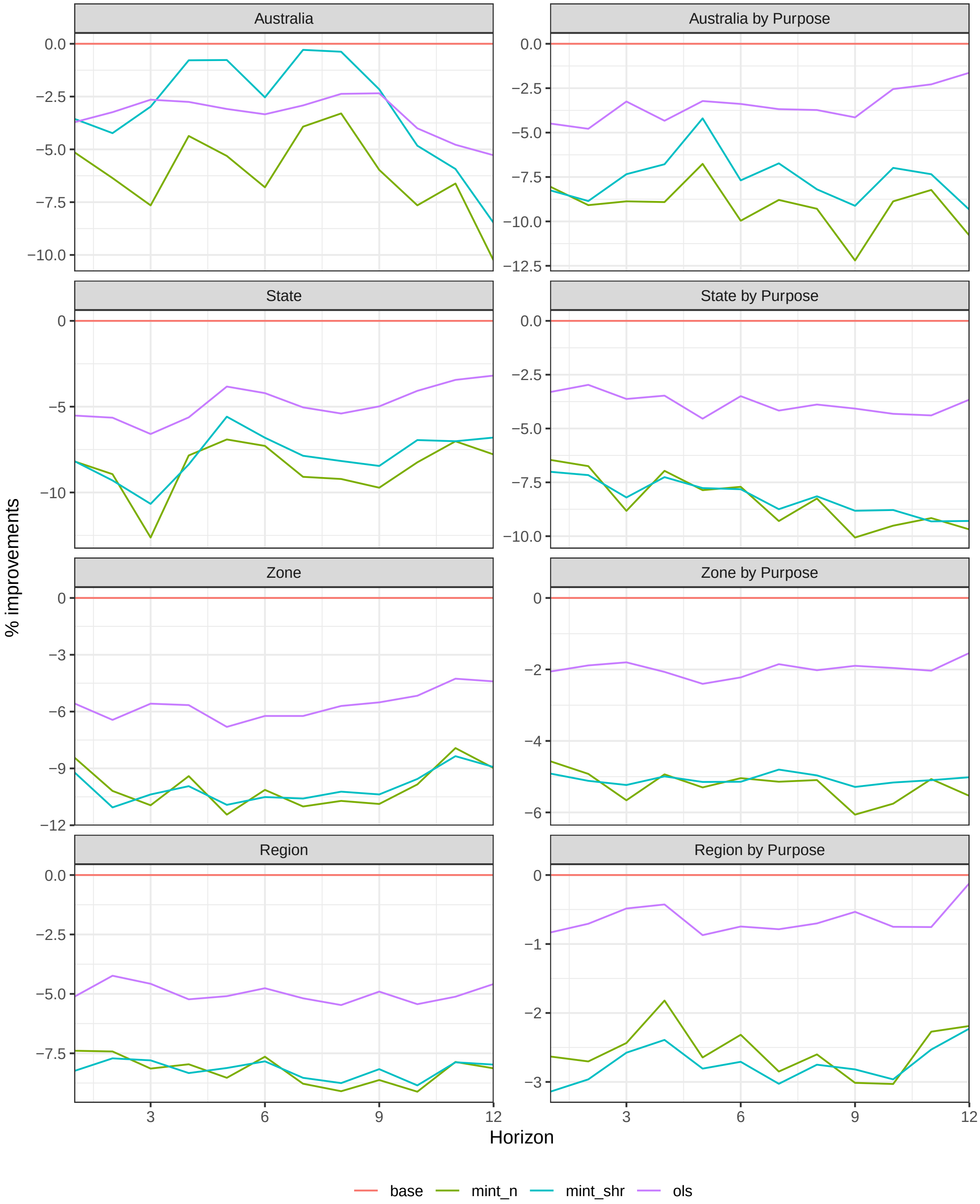

Figure 10: Relative improvement of the MSE of reconciled forecasts over the base forecasts for the Australian domestic tourism flows, excluding the period after 2016

6 Timeline & Milestones

7 Appendix

7.1 Simulation Settings: Supplementary Figures

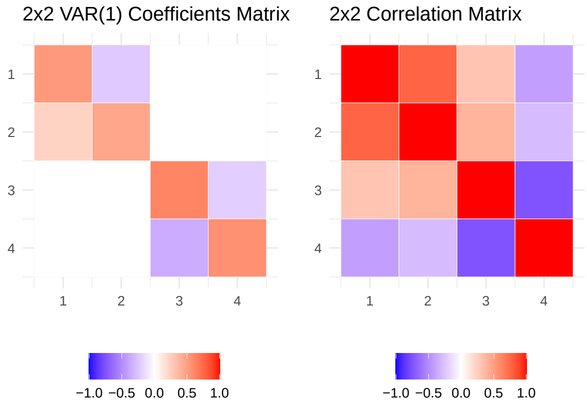

Figure 11: Heatmaps of the VAR(1) coefficient matrix and correlation matrix for the 2x2 structure.

7.2 Algorithm: NOVELIST cross-validation for optimal threshold δ^*

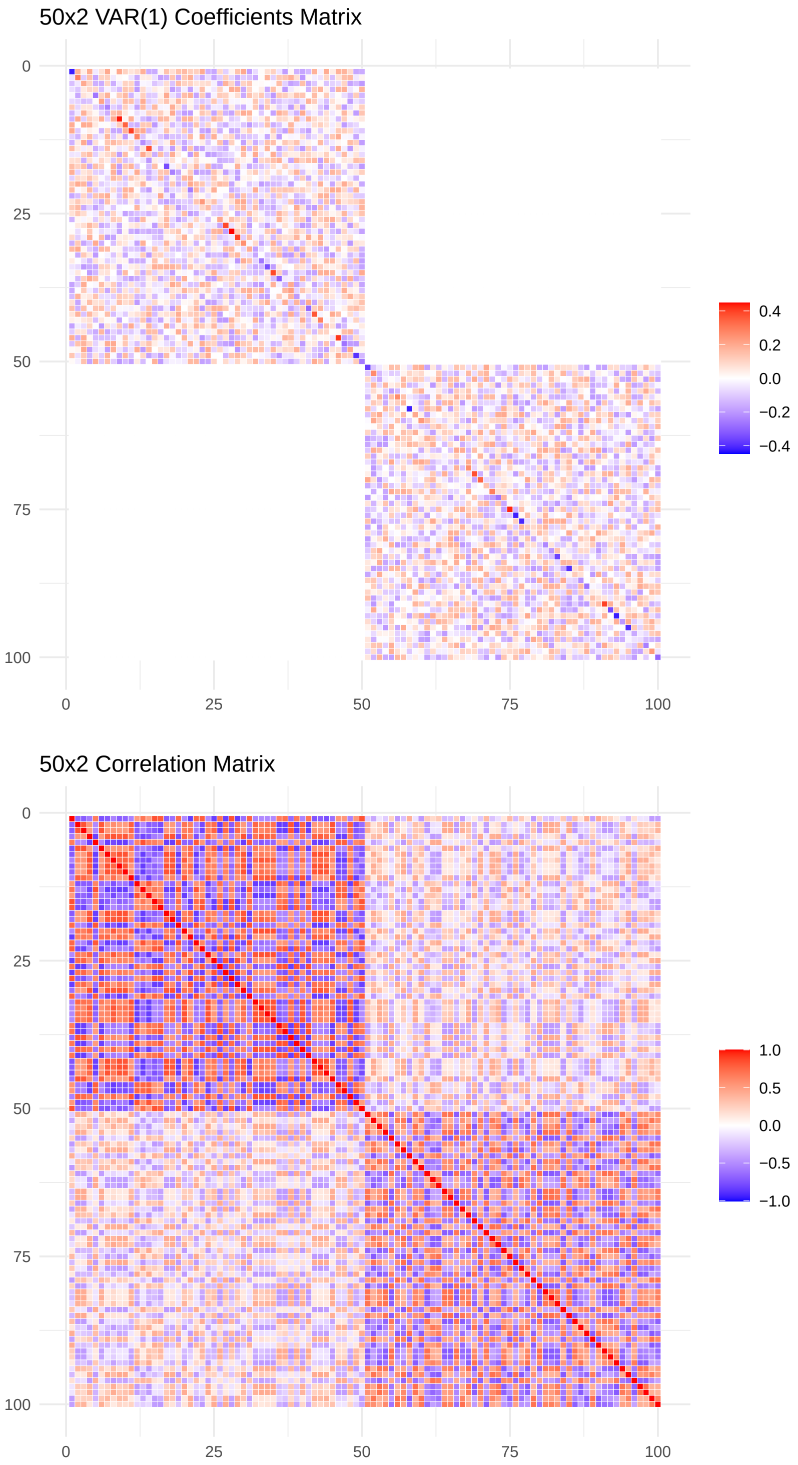

Figure 12: Heatmaps of the VAR(1) coefficient matrix and correlation matrix for the 6×6 structure.

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 $\Delta$, 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 $\hat{\delta}^*$
```

References

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

Bandara, K., Hyndman, R. J., & Bergmeir, C. (2022). MSTL: A seasonal-trend decomposition algorithm for time series with multiple seasonal patterns. *arXiv [Stat.AP]*. http://arxiv.org/abs/2107.13462

Fan, J., Liao, Y., & Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *J. R. Stat. Soc. Series B Stat. Methodol.*, *75*(4), 603–680. https://doi.org/10.1111/rssb.12016

Hardin, J., Garcia, S. R., & Golan, D. (2013). A method for generating realistic correlation matrices. *Ann. Appl. Stat.*, *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 (Madr.)*, *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. *Comput. Stat. Data Anal.*, *55*(9), 2579–2589. https://doi.org/10.1016/j.csda.2011.03.006

Hyndman, R. J., Lee, A. J., & Wang, E. (2016). Fast computation of reconciled forecasts for hierarchical and grouped time series. *Comput. Stat. Data Anal.*, *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. *J. Empir. Finance*, *10*(5), 603–621. https://doi.org/10.1016/s0927-5398(03)00007-0

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

Schäfer, J., & Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.*, *4*(1), Article32. https://doi.org/10.2202/1544-6115.1175

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

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