Enhancing Forecast Reconciliation: A Study of Alternative Covariance Estimators

1 Objectives

When forecasting sales for items in a cafe, such as matcha latte and mocha, the forecasts for each of these drinks (110 matcha lattes and 90 mochas) are often not consistent with the overall sales forecast for the cafe (180 drinks). This is a problem of forecasting hierarchical time series, where the individual forecasts do not satisfy the linear constraints of different levels of aggregation. Forecast reconciliation comes in to solve this problem, where the individual forecasts are adjusted to satisfy the given constraints. Among the various reconciliation methods, the **MinT** (Minimum Trace) is considered the optimal approach. However, this method requires an 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 neglect the prominent 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.

2 Background

In time series forecasting, aggregation occurs in a variety of settings. A concrete example of a hierarchy would be electricity demand forecasting, where the national demand is the sum of the demands for each state, and demand for each state comes from many regions within the states. Forecasting national tourism or Gross Domestic Product (GDP) is another example of hierarchical/grouped time series. The impact of methods for forecasting hierarchical time series has not been limited to academia, with industry also showing a strong interest. Many companies have adopted these methods in practice, including Amazon, the International Monetary Fund, IBM, SAP, and more. (Athanasopolous, 2024)

The hierarchical structure can be represented as a tree, as shown in figure 1. The top level of the tree represents the total forecast, while the lower levels represent the individual forecasts.

When there are attributes of interest that are crossed, such as the forecast for electricity demand can be broken down by usage purposes (e.g., residential and commercial), it will become a grouped time series.

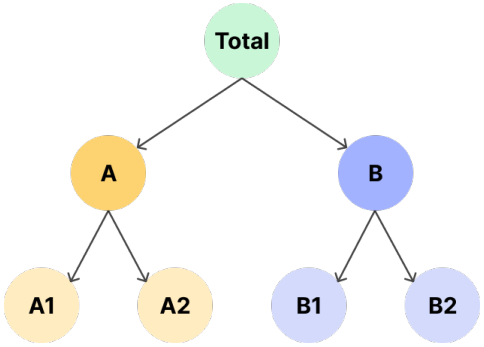

Figure 1: Diagram of 2-level hierarchical tree structure

Both of these structure can be represented using matrix algebra:

$$\boldsymbol{y}_t = \boldsymbol{S}\boldsymbol{b}_t, \tag{1}$$

where $\boldsymbol{S}$ is a summing matrix of order $n \times n_b$ which aggregates the bottom-level series $\boldsymbol{b}_t$ (n_b-vector) to the series at aggregation levels above. The n-vector $\boldsymbol{y}_t$ contains all observations at time t.

The example summing matrix $\boldsymbol{S}$ for the tree structure in Figure 1 is:

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

The MinT method is a popular choice for estimating $\boldsymbol{G}_h$, as it minimizes the trace of the covariance matrix of the forecast errors, which is a measure of the uncertainty in the forecasts. The MinT method requires an estimate of the covariance matrix of the base forecast errors, which is typically obtained using a shrinkage estimator. However, this approach has limitations, as it may not fully capture the structure of the data.

3 Methodology

If we use Equation 1 to forecast the time series, we would not be utilising all the information in the data, since we only forecast the bottom-level series $\boldsymbol{b}_t$ then aggregate to the higher-level. Thus, Hyndman et al. (2011) showed that existing methods could be expressed as:

$$\tilde{\boldsymbol{y}}_h = \boldsymbol{S}\boldsymbol{G}_h\hat{\boldsymbol{y}}_h, \tag{2}$$

for a suitable $n_b \times n$ matrix $\boldsymbol{G}_h$ (we can drop the subscript h when G does not depend on the forecast horizon h). $\boldsymbol{G}_h$ maps the base forecasts of all levels $\hat{\boldsymbol{y}}_h$ down into the bottom series, which is then aggregated to the higher levels by $\boldsymbol{S}$. Note that any method may have been used to produce the base forecasts.

From this we can see the importance of the matrix $\boldsymbol{G}_h$, and the choice of it determines the performance of reconciled forecasts $\tilde{\boldsymbol{y}}_h$. Methods are developed to estimate $\boldsymbol{G}_h$, including the OLS and WLS, proposed by Hyndman et al. (2011) and Hyndman et al. (2016) respectively.

3.1 Mininum Trace (MinT) Reconciliation

Wickramasuriya et al. (2019) reframed the problem by taking an optimisation approach rather than the regression. They formulated the problem as minimising the variances of all reconciled forecasts from Equation 2, which happens to be equivalent to the trace of the covariance matrix (sum of the diagonal elements). This is known as the Minimum Trace (MinT) reconciliation method. The MinT solution is given by

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

and $\boldsymbol{W}_h$ is the covariance matrix of the h-step-ahead base forecast errors.

The MinT approach is an algebraical generalisation of the GLS, and the OLS and WLS methods are special cases of MinT when $\boldsymbol{W}_h$ is a diagonal or identity matrix, respectively. However, the MinT solution hinges on a reliable estimate of $\boldsymbol{W}_h$, which is challenging to estimate in high-dimensional setting. Therefore, we need alternative covariance estimators.

3.2 Alternative Covariance Estimators

We reconstruct estimator of $\boldsymbol{W}_h$ as $\hat{\boldsymbol{W}}_h = k_h\, g(\hat{\boldsymbol{W}}_1)$, and $g(.)$ is an estimator function of the unbiased sample covariance of in-sample one-step-ahead base forecast errors $\hat{\boldsymbol{W}}_1 = \frac{1}{T}\sum_{t=1}^{T}\hat{\boldsymbol{e}}_{t|t-1}\hat{\boldsymbol{e}}'_{t|t-1}$

(a) Shrinkage

The proposed MinT approach by Wickramasuriya et al. (2019) uses the shrinkage estimator from Schäfer and Strimmer (2005). The shrinkage estimator is given by:

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

$\hat{\boldsymbol{W}}_{1,D}$ is a diagonal matrix comprising the diagonal entries of $\hat{\boldsymbol{W}}_1$. This approach will shrink the covariance matrix $\hat{\boldsymbol{W}}_1$ towards its diagonal version, meaning the off-diagonal elements are shrunk towards zero while the diagonal ones remain unchanged.

Schäfer and Strimmer (2005) also proposed an optimal shrinkage intensity parameter λ_D for this setting, assuming the variances are constant:

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

where $\hat{r}_{ij}$ is the ijth element of $\hat{\boldsymbol{R}}_1$, the 1-step-ahead sample correlation matrix (obtained from $\hat{\boldsymbol{W}}_1$) to shrink it toward an identity matrix

The optimal lambda is obtained by minimising the $MSE(\hat{\boldsymbol{W}}_1) = Bias(\hat{\boldsymbol{W}}_1)^2 + Var(\hat{\boldsymbol{W}}_1)$. More specifically, we trade the unbiasedness of the sample covariance matrix for a lower variance. The objective function itself does not take into account any possible principal components structure in the data, and is not flexible enough since it shrinks all off-diagonal elements equally towards zeros.

(b) NOVELIST

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 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, offering more flexibility. The NOVELIST estimator is given by:

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

By convenient setting, we rewrite it as in correlation matrix:

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

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

$$\hat{r}^\delta_{1,ij} = \text{sign}(\hat{r}_{1,ij})\,(|\hat{r}_{1,ij}| - \delta)_+.$$

For a given threshold δ, Huang & Fryzlewicz derived the optimal shrinkage intensity parameter $\lambda(\delta)$ using Ledoit-Wolf's lemma (Ledoit and Wolf, 2003). It can be computed as:

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

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, for each rolling-window set, loop through a range of threshold values, compute the corresponding $\hat{\lambda}(\delta)$ and $\hat{\boldsymbol{R}}_1^N$. Then we select the one that minimises the average out-of-sample reconciled forecast errors. Although it is not required to fit forecasting models multiple time, the cross-validation procedure is still computationally expensive as it computes the NOVELIST estimator and perform reconciliation for each threshold value. The formal algorithm is given in Figure 2, in the Section 8 Appendix.

(c) POET

The POET (Principal Orthogonal complEment Thresholding) estimator, proposed by Fan et al. (2013), is another "sparse" + "non-sparse" covariance estimator. It 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 eigenvalue and eigenvector of the sample covariance matrix. $T(.)$ is the thresholding function, which can be either soft-thresholding, hard-thresholding, or others.

(d) NOVELIST with PC-adjusted

This approach is similar to the POET estimator. The difference is that POET apply thresholding to the orthogonal complement part $\hat{\boldsymbol{R}}_{1,K}$, meanwhile we apply NOVELIST. It can be formulated as:

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

and $\hat{\boldsymbol{R}}_{1,K}^N$ is the NOVELIST estimator applied to the orthogonal complement part $\hat{\boldsymbol{R}}_{1,K}$.

4 Experimental Design

The experimental design is to simulate a hierarchical time series data set, then apply the MinT reconciliation method with different covariance estimators. The data set will be split into training and test sets. In case of cross-validation, the training set will be further split into training and validation sets.

```
--------- put these into appendix -----------
```

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

$$\boldsymbol{y}_t = \boldsymbol{A}\boldsymbol{y}_{t-1} + \boldsymbol{e}_t,$$

where $\boldsymbol{A}$ is a $p \times p$ block diagonal matrix of autoregressive coefficients $\boldsymbol{A} = diag(\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m)$, with each $\boldsymbol{A}_i$ being a $p_i \times p_i$ 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 better performed than the usual base forecasts.

The model is added with a Gaussian innovation process $\boldsymbol{e}_t$, with covariance matrix Σ. The covariance matrix Σ is generated specifically in the following way:

1. A compound symmetric correlation matrix is used for each block of size p_i in $\boldsymbol{A}_i$, where the coefficients are sampled from a uniform distribution.
2. The correlations between different blocks are imposed using the Algorithm 1 in Hardin, Garcia & Golan (2013).
3. The covariance matrix Σ is then constructed by uniform sampling of standard deviations for all p series.

We have an option to randomly flip the signs of the covariance elements, which will create a more realistic structure in the innovation process. This is also to simulate the real-world scenario where the observed covariance matrix is not necessarily positive definite.

```
----------------------------------------
```

- Put some graphs here

5 Timeline & Milestones

6 Expected Contributions

7 References

8 Appendix

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

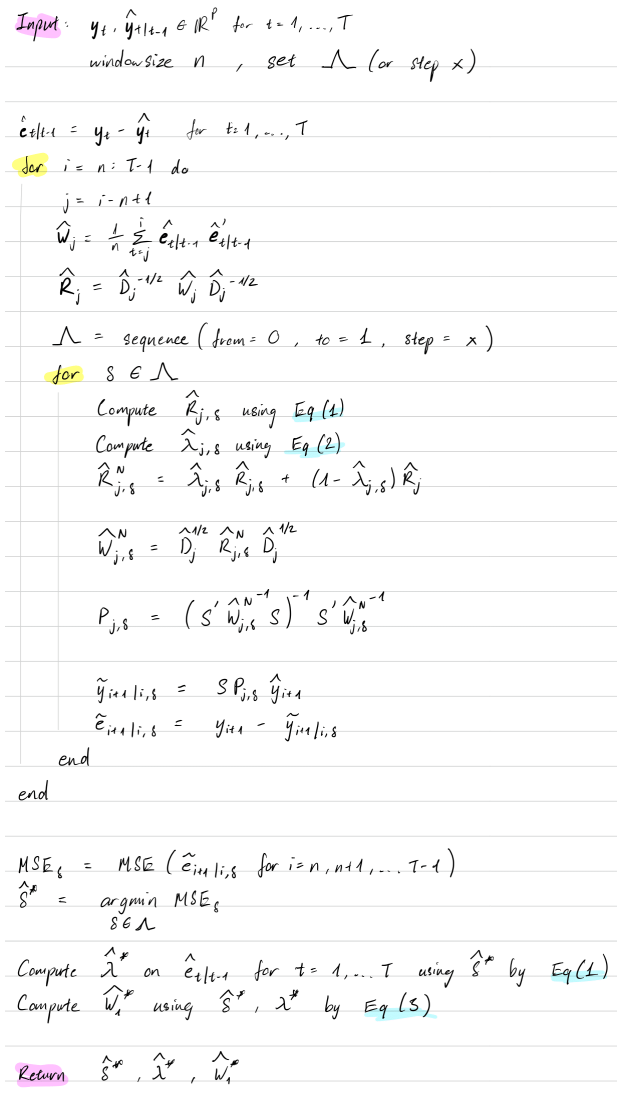

Figure 2: NOVELIST cross-validation procedure

aaa