

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.

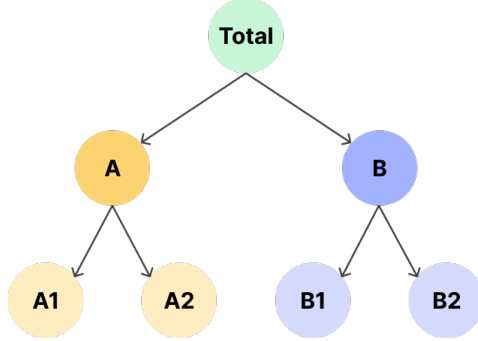


Figure 1: Diagram of 2-level hierarchical tree structure

Both of these structure can be represented using matrix algebra:

$$\mathbf{y}_t = \mathbf{S}\mathbf{b}_t, \quad (1)$$

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

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

$$\mathbf{S} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ \mathbf{I}_n \end{bmatrix},$$

The MinT method is a popular choice for estimating \mathbf{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 \mathbf{b}_t then aggregate to the higher-level. Thus, Hyndman et al. (2011) showed that existing methods could be expressed as:

$$\tilde{\mathbf{y}}_h = \mathbf{S}\mathbf{G}_h\hat{\mathbf{y}}_h, \quad (2)$$

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

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

3.1 Minimum 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

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

and \mathbf{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 \mathbf{W}_h is a diagonal or identity matrix, respectively. However, the MinT solution hinges on a reliable estimate of \mathbf{W}_h . In high-dimensional setting, the usual sample covariance matrix is unstable, thus we need alternative covariance estimators.

3.2 Alternative Covariance Estimators

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:

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

and $\hat{\mathbf{W}}_1 = \frac{1}{T} \sum_{t=1}^T \hat{\mathbf{e}}_{t|t-1} \hat{\mathbf{e}}'_{t|t-1}$ is the unbiased sample covariance of in-sample one-step-ahead base forecast errors.

$\hat{\mathbf{W}}_{1,D}$ is a diagonal matrix comprising the diagonal entries of $\hat{\mathbf{W}}_1$. This approach will shrink the covariance matrix $\hat{\mathbf{W}}_1$ 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 ij th element of $\hat{\mathbf{R}}_1$, the 1-step-ahead sample correlation matrix (obtained from $\hat{\mathbf{W}}_1$) to shrink it toward an identity matrix

- Novelist
- POET
- Novelist with PC-adjusted

4 Experimental Design

5 Timeline & Milestones

6 Expected Contributions

7 REferences