Enhancing Forecast Reconciliation: A Study of Alternative Covariance Estimators

1 Objectives

Globally, Starbucks manages thousands of outlets across dozens of countries, each projecting sales for dozens of beverages. Yet when regional, national, and brand-wide forecasts are generated, discrepancies of millions of dollars emerge simply because individual outlet forecasts fail to respect the hierarchical nature of the data. Forecast reconciliation comes in to solve this huge problem, where the individual forecasts are adjusted to satisfy the aggregation constraints. Among the various reconciliation methods, **MinT** (Minimum Trace) is considered the optimal approach, 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 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 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 (Athanasopoulos et al., 2024).

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 electricity demand at any aggregation level (national, state, or regional) is also considered by usage purposes (e.g., residential, commercial), the structure is described as a grouped time series (illustrated in Figure 2).

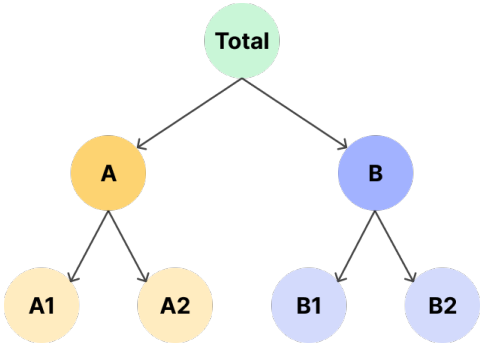

Figure 1: Diagram of 2-level hierarchical tree structure

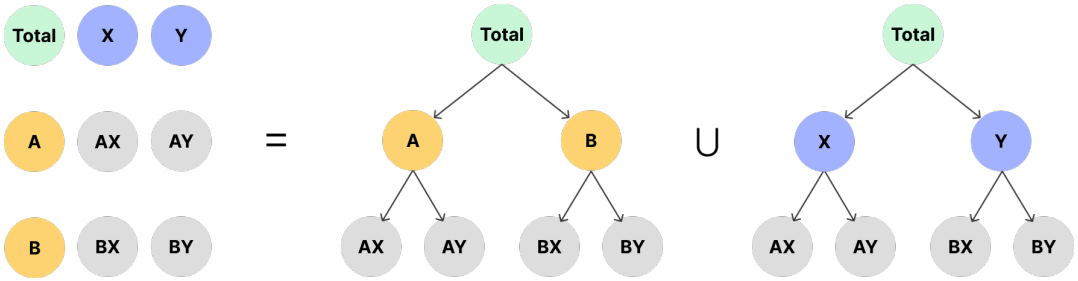

Figure 2: Diagram of 2-level grouped structure, which can be considered as the union of two hierarchical trees with common top and bottom level series

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 $\boldsymbol{b}_t$ (n_b-vector) to the series at aggregation levels above. The n-vector $\boldsymbol{y}_t$ contains all observations at time t.

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

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

3 Methodology

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

$$\tilde{\boldsymbol{y}}_h = \boldsymbol{S}\hat{\boldsymbol{b}}_h. \tag{1}$$

We refer to $\tilde{\boldsymbol{y}}_h$ 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 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:

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

for a suitable $n_b \times n$ matrix $\boldsymbol{G}$. $\boldsymbol{G}$ maps the base forecasts of all levels $\hat{\boldsymbol{y}}_h$ 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 $\tilde{\boldsymbol{y}}_h$. Methods are developed to estimate $\boldsymbol{G}$, the first attempts at least squares forecast reconciliation were made by Hyndman et al. (2011).

They proposed the optimal $\boldsymbol{G}$ based on the regression model:

$$\hat{\boldsymbol{y}}_h = \boldsymbol{S}\boldsymbol{\beta}_h + \boldsymbol{\epsilon}_h,$$

where $\boldsymbol{\epsilon}_h = \tilde{\boldsymbol{y}}_h - \hat{\boldsymbol{y}}_h$ is the coherency error with variance $\boldsymbol{V}_h$. This led to the GLS solution: $\boldsymbol{G} = (\boldsymbol{S}'\boldsymbol{V}_h^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{V}_h^{-1}$.

However, the covariance matrix $\boldsymbol{V}_h$ is unknown (and later shown by Wickramasuriya et al. (2019) to be unidentifiable), and replaced by an identity matrix. The method then collapses into an OLS solution: $\boldsymbol{G} = (\boldsymbol{S}'\boldsymbol{S})^{-1}\boldsymbol{S}'$.

An obvious drawback of the OLS solution is that it weights all series in any level equally, regardless of its scale or forecast error variance. The issue prompted Hyndman et al. (2016) to propose a WLS solution, where the series are weighted by the inverse variances of the base forecast errors. The WLS solution is $\boldsymbol{G} = (\boldsymbol{S}'\boldsymbol{\Lambda}_h^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{\Lambda}_h^{-1}$. $\boldsymbol{\Lambda}_h = \text{diag}(\boldsymbol{W}_h)$ and $\boldsymbol{W}_h = Var(\boldsymbol{y}_h - \hat{\boldsymbol{y}}_h)$.

3.1 Minimum Trace (MinT) Reconciliation

Wickramasuriya et al. (2019) reframed the problem by taking an optimisation approach. They formulated the problem as minimising the variances of all reconciled forecasts from Equation 2. They showed that this is equivalent to minimising the trace of the base forecast error covariance matrix (sum of the diagonal elements - the variances). This is known as the Minimum Trace (MinT) reconciliation method. The MinT solution is given by

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

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

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 and a diagonal matrix, respectively.

The MinT solution hinges on a reliable, positive-definite estimate of $\boldsymbol{W}_h$, which is challenging to estimate in high-dimensional setting. Therefore, we will be exploring alternative covariance estimators.

3.2 Alternative Covariance Estimators

We reconstruct the estimator of $\boldsymbol{W}_h$ as $\hat{\boldsymbol{W}}_h = k_h\, g(\hat{\boldsymbol{W}}_1)$, where $k_h > 0$. The function $g(.)$ is an estimator of the unbiased sample covariance matrix of the in-sample one-step-ahead base forecast errors $\hat{\boldsymbol{W}}_1 = \frac{1}{T}\sum_{t=1}^{T}\hat{\boldsymbol{e}}_t\hat{\boldsymbol{e}}_t'$, and T is the length of series' time dimension.

(a) Shrinkage

The proposed MinT approach by Wickramasuriya et al. (2019) uses the shrinkage estimator from Schäfer & Strimmer (2005). It guarantees positive definiteness and variance reduction for the covariance matrix, especially when the total number of series $n > T$. The shrinkage estimator is 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 of $\hat{\boldsymbol{W}}_1$. 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 λ_D is obtained by minimising the mean squared error of $\hat{\boldsymbol{W}}_1$: $MSE(\hat{\boldsymbol{W}}_1) = Bias(\hat{\boldsymbol{W}}_1)^2 + Var(\hat{\boldsymbol{W}}_1)$. 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.

- Screeplot of eigenvalues

(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 introduces adaptive sparsity, retaining strong correlations while discarding weak, noisy correlations. NOVELIST offers more flexibility than the shrinkage estimator, 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:

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

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). It can be computed as:

- @Huang2019-ua mentioned Ledoit-Wolf's lemma

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

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 formal algorithm is given in the Section 7.1 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.

(c) 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 into account, and is appealing when there are common drivers in the time series within the hierarchy.

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.

(d) 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 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. The data generating process is described in Section 7.2 Appendix.

- Put some graphs here

5 Timeline & Milestones

Table 1: Project Timeline and Key Milestones

Period	Task	Deliverable
March - April	Literature review, methods implementation (shrinkage, NOVELIST) + simulation framework	Main codes + paper draft
May	Simulation + NOVELIST assessment	Results: NOVELIST
June	Experiments with POET and PC-adjusted NOVELIST	Results: POET & PC-adjusted NOVELIST
July	Exploration of other high-dimensional covariance estimators	Results: other estimators
August - September	Real-world application + evaluation	Application completed
October	Final writing + submission	Thesis manuscript

6 Expected Contributions

7 Appendix

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

Algorithm 1 Cross-validation procedure

Input: Observations and fitted values $\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t \in \mathbb{R}^n$ for $t = 1, \ldots, T$, set of threshold candidates Δ, window size v.
$\hat{\boldsymbol{e}}_t = \boldsymbol{y}_t - \hat{\boldsymbol{y}}_t$ for $t = 1, \ldots, T$
for $i = v : T-1$ do **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 $\hat{\delta}^*$

7.2 Data generating progress design

The designed 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 better performed than the usual base forecasts.

The model is added with a Gaussian innovation process $\boldsymbol{\epsilon}_t$, 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 $n_{b,i}$ in $\boldsymbol{A}_i$, where the coefficients are sampled from a uniform distribution between 0 and 1.
2. The correlations between different blocks are imposed using the Algorithm 1 in Hardin et al. (2013).
3. The covariance matrix Σ is then constructed by uniform sampling of standard deviations, in a range of $[\sqrt{2}, \sqrt{6}]$, for all n_b 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 can be done by pre- and post-multiplying Σ by a random diagonal matrix V with entries sampled from $\{-1, 1\}$, yielding $\Sigma' = V\Sigma V$.

References

Athanasopoulos, G., Hyndman, R. J., Kourentzes, N., & Panagiotelis, A. (2024). *Forecast reconciliation: A review. 40*(2), 430–456.

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.

Hardin, J., Garcia, S. R., & Golan, D. (2013). A method for generating realistic correlation matrices. *Ann. Appl. Stat.*, *7*(3), 1733–1762.

Higham, N. (2002). Computing the nearest correlation matrix—a problem from finance. *Ima Journal of Numerical Analysis*, *22*, 329–343.

Huang, N., & Fryzlewicz, P. (2019). NOVELIST estimator of large correlation and covariance matrices and their inverses. *Test (Madr.)*, *28*(3), 694–727.

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.

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.

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.

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.

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.