MONASH University

MONASH BUSINESS SCHOOL

ISSN 1440-771X

Department of Econometrics and Business Statistics

http://monash.edu/business/ebs/research/publications

# Optimal forecast reconciliation with time series selection

Xiaoqian Wang, Rob J Hyndman, Shanika L Wickramasuriya

September 2023

Working Paper no/yr

# Optimal forecast reconciliation with time series selection

**Xiaoqian Wang**
Monash University, VIC 3800, Australia
Email: xiaoqian.wang@monash.edu
Corresponding author

**Rob J Hyndman**
Monash University, VIC 3800, Australia
Email: rob.hyndman@monash.edu

**Shanika L Wickramasuriya**
Monash University, VIC 3145, Australia
Email: shanika.wickramasuriya@monash.edu

25 September 2023

# Optimal forecast reconciliation with time series selection

**Abstract**

Abstract

**Keywords:** Keyword 1, Keyword 2

## 1 Introduction

Hierarchical time series and forecast reconciliation. Post-processing.

Single-level approaches, least squares-based reconciliation approaches, geometric intuition, other extensions with constraints.

However… Two issues. The choice of W can have significant effect on the quality of the reconciled forecasts. Some time series perform poorly.

In this paper, our focus will be on… selection. Other entries will be adjusted accordingly.

The remainder of the paper is structured as follows.

## 2 Preliminaries

### 2.1 Notation

We denote the set $\{1, \ldots, k\}$ by $[k]$ for any non-negative integer $k$. A *hierarchical time series* can be considered as an $n$-dimensional multivariate time series, $\{\boldsymbol{y}_t, t \in [T]\}$, that adheres to known linear constraints. Let $\boldsymbol{y}_t \in \mathbb{R}^n$ be a vector comprising observations of all time series in the hierarchy at time $t$, and $\boldsymbol{b}_t \in \mathbb{R}^{n_b}$ be a vector comprising observations of all bottom-level time series at time $t$. The full hierarchy at time $t$ can be written as

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

where $\boldsymbol{S}$ is an $n \times n_b$ *summing matrix* that shows aggregation constraints present in the structure. We can write the summing matrix as $\boldsymbol{S} = \begin{bmatrix} \boldsymbol{A} \\ \boldsymbol{I}_{n_b} \end{bmatrix}$, where $\boldsymbol{A}$ is an $n_a \times n_b$ *aggregation matrix* with $n = n_a + n_b$, and $\boldsymbol{I}_{n_b}$ is an $n_b$-dimensional identity matrix.

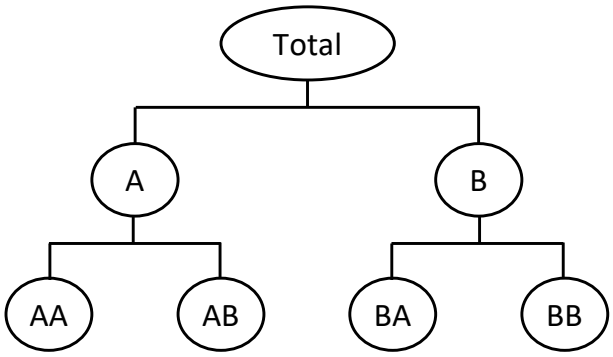


**Figure 1:** *An example of a two-level hierarchical time series.*

To clarify these notations, consider the example of the hierarchy in Figure 1. For this two-level hierarchy, $n = 7$, $n_b = 4$, $n_a = 3$, $\boldsymbol{y}_t = [y_{\text{Total},t}, y_{\text{A},t}, y_{\text{B},t}, y_{\text{AA},t}, y_{\text{AB},t}, y_{\text{BA},t}, y_{\text{BB},t}]'$, $\boldsymbol{b}_t = [y_{\text{AA},t}, y_{\text{AB},t}, y_{\text{BA},t}, y_{\text{BB},t}]'$, and

$$\boldsymbol{S} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ & \boldsymbol{I}_4 & & \end{bmatrix}.$$

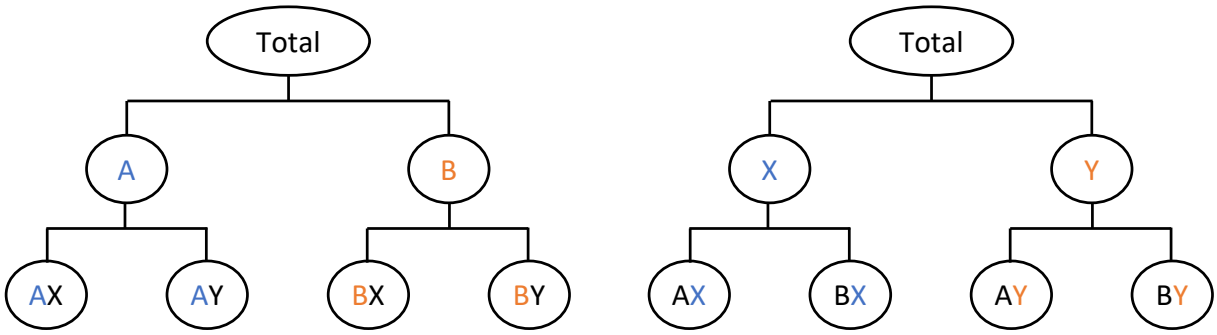


**Figure 2:** *An example of a two level grouped time series.*

When data structure does not naturally disaggregate in a unique hierarchical manner, we can combine these hierarchical structures to form a *grouped time series*. Thus, grouped time series can also be considered as hierarchical time series with more than one grouping structure. Figure 2 shows an example of a two level grouped time series with two alternative aggregation structures. For this example, $n = 9$, $n_b = 4$, $n_a = 5$, $\boldsymbol{y}_t = [y_{\text{Total},t}, y_{\text{A},t}, y_{\text{B},t}, y_{\text{X},t}, y_{\text{Y},t}, y_{\text{AX},t}, y_{\text{AY},t}, y_{\text{BX},t}, y_{\text{BY},t}]'$, $\boldsymbol{b}_t = [y_{\text{AX},t}, y_{\text{AY},t}, y_{\text{BX},t}, y_{\text{BY},t}]'$, and

$$\boldsymbol{S} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ & \boldsymbol{I}_4 & \end{bmatrix}.$$

## 2.2 Linear forecast reconciliation

Let $\hat{\boldsymbol{y}}_{T+h|T} \in \mathbb{R}^n$ be a vector of $h$-step-ahead *base forecasts* for all time series in the hierarchy, given observations up to time $T$, and stacked in the same order as $\boldsymbol{y}_t$. We can use any method to generate these forecasts, but In general they will not add up especially when we forecast each series independently.

When forecasting hierarchical time series, we expect the forecasts to be *coherent* (i.e., aggregation constraints are satisfied). Let $\tilde{\boldsymbol{y}}_{T+h|T} \in \mathbb{R}^n$ denote a vector of $h$-step-ahead *reconciled forecasts* which are coherent by construction, $\psi$ a mapping that reconciles base forecasts, $\hat{\boldsymbol{y}}_{T+h|T}$. Then we have *forecast reconciliation* $\tilde{\boldsymbol{y}}_{T+h|T} = \psi(\hat{\boldsymbol{y}}_{T+h|T})$, which is essentially a post-processing method. In this paper, we focus on linear forecast reconciliation given by

$$\tilde{\boldsymbol{y}}_{T+h|T} = \boldsymbol{S}\boldsymbol{G}_h\hat{\boldsymbol{y}}_{T+h|T},$$

where

- $\boldsymbol{G}_h$ is an $n_b \times n$ weighting matrix that maps the base forecasts into the bottom level. In other words, it combines all base forecasts to form reconciled forecasts for bottom-level series.
- $\boldsymbol{S}$ is an $n \times n_b$ summing matrix that sums up bottom-level reconciled forecasts to produce coherent forecasts of all levels. It identifies the linear constraints involved in the hierarchy.

### 2.2.1 Minimum trace reconciliation

Let the $h$-step-ahead *base forecast errors* be defined as $\hat{\boldsymbol{e}}_{T+h|T} = \boldsymbol{y}_{T+h} - \hat{\boldsymbol{y}}_{T+h|T}$, and the $h$-step-ahead *reconciled forecast errors* be defined as $\tilde{\boldsymbol{e}}_{T+h|T} = \boldsymbol{y}_{T+h} - \tilde{\boldsymbol{y}}_{T+h|T}$. Wickramasuriya, Athanasopoulos & Hyndman (2019) formulated a linear reconciliation problem as minimizing the trace (MinT) of the $h$-step-ahead covariance matrix of the reconciled forecast errors, $\text{Var}(\tilde{\boldsymbol{e}}_{T+h|T})$.

Under the assumption of unbiasedness, the unique solution of the minimization problem is given by

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

where $\boldsymbol{W}_h$ is the positive definite covariance matrix of the $h$-step-ahead base forecast errors, $\text{Var}(\hat{\boldsymbol{e}}_{T+h|T})$.

The trace minimization problem can be reformulated as a least squares problem with linear constraints given by

$$\begin{aligned} \min_{\tilde{\boldsymbol{y}}_{T+h|T}} \quad & \frac{1}{2}(\hat{\boldsymbol{y}}_{T+h|T} - \tilde{\boldsymbol{y}}_{T+h|T})'\boldsymbol{W}_h^{-1}(\hat{\boldsymbol{y}}_{T+h|T} - \tilde{\boldsymbol{y}}_{T+h|T}) \\ \text{s.t.} \quad & \tilde{\boldsymbol{y}}_{T+h|T} = \boldsymbol{S}\tilde{\boldsymbol{b}}_{T+h|T}, \end{aligned} \tag{2}$$

where $\tilde{\boldsymbol{b}}_{T+h|T} \in \mathbb{R}^{n_b}$ is the vector comprising $h$-step-ahead bottom-level reconciled forecasts, made at time $T$. Focusing on $\boldsymbol{W}_h$, the intuitive behind the MinT reconciliation is that the larger the estimated variance of the base forecast errors, the larger the range of adjustments permitted for forecast reconciliation.

It's challenging to estimate $\boldsymbol{W}_h$, especially for $h > 1$. Assuming that $\boldsymbol{W}_h = k_h\boldsymbol{W}_1$, $\forall h$, where $k_h > 0$, the MinT solution of $\boldsymbol{G}$ does not change with the forecast horizon, $h$. Hence, we will drop the subscript $h$ for the ease of exposition. The most popularly used candidate estimators for $\boldsymbol{W}$ in the forecast reconciliation literature are listed as follows.

1. $\boldsymbol{W}_{\text{OLS}} = \boldsymbol{I}$ is the *OLS estimator* proposed by Hyndman et al. (2011), assuming that the base forecast errors are uncorrelated and equivariant. In what follows, we denote this as **OLS**.
2. $\boldsymbol{W}_{\text{WLSs}} = \text{diag}(\boldsymbol{S}\boldsymbol{1})$ is the *WLS estimator applying structural scaling* proposed by Athanasopoulos et al. (2017). This estimator depends only on the aggregation structure of the hierarchy. It assumes that the variance of each bottom-level base forecast error is equivalent and uncorrelated between nodes. We denote this method as **WLSs**.
3. $\boldsymbol{W}_{\text{WLSv}} = \text{diag}(\hat{\boldsymbol{W}}_1)$ is the *WLS estimator applying variance scaling* proposed by Hyndman, Lee & Wang (2016), where $\hat{\boldsymbol{W}}_1$ denotes the unbiased covariance estimator based on the in-sample one-step-ahead base forecast errors (i.e., residuals). In the results that follow, we denote this as **WLSv**.

4. $\boldsymbol{W}_{\text{MinT}} = \hat{\boldsymbol{W}}_1$ is referred to as the *MinT estimator* based on the sample covariance matrix proposed by Wickramasuriya, Athanasopoulos & Hyndman (2019). We denote this method as **MinT** in the results that follow.
5. $\boldsymbol{W}_{\text{MinTs}} = \lambda\,\text{diag}(\hat{\boldsymbol{W}}_1) + (1-\lambda)\hat{\boldsymbol{W}}_1$ is the *MinT shrinkage estimator* suggested by Wickramasuriya, Athanasopoulos & Hyndman (2019), in which off-diagonal elements of $\hat{\boldsymbol{W}}_1$ are shrunk toward zero. We refer to this method as **MinTs**.

It's hard to say which estimator for $\boldsymbol{W}$ works better. Pritularga, Svetunkov & Kourentzes (2021) demonstrated that the performance of forecast reconciliation is affected by two sources of uncertainties, i.e., the base forecast uncertainty and the reconciliation weight uncertainty. Recall that the uncertainty in the MinT solution in Equation 1 is introduced by the uncertainty in the reconciliation weighting matrix as the summing matrix is fixed for a certain hierarchy. This indicates that OLS and WLSs estimators for $\boldsymbol{W}$ may lead to less volatile reconcliation performance compared to WLSv, MinT, and MinTs estimators. Panagiotelis et al. (2021) provided a geometric intuition for reconciliation and showed that, when considering the Euclidean distance loss function, OLS reconciliation yields results that are at least as favorable as the base forecasts, whereas MinT reconciliation performs poorly relative to the base forecasts. However, when considering the mean squared reconciled forecast error, Wickramasuriya (2021) indicated that MinT reconciliation is better than OLS reconciliation. Therefore, which estimator for $\boldsymbol{W}$ to use hinges on the specific hierarchical time series of interest, the targeted level or series, and the selected loss function.

### 2.2.2 Relaxation of the unbiasedness assumptions

Both Hyndman et al. (2011) and Wickramasuriya, Athanasopoulos & Hyndman (2019) impose two unbiasedness conditions, i.e., the base forecasts and the reconciled forecasts are unbiased. Ben Taieb & Koo (2019) proposed a reconciliation method relaxing the assumption of unbiasedness. Specifically, by expanding the training window forward by one observation until $T-h$, they formulated the reconciliation problem as a regularized empirical risk minimization (RERM) problem given by

$$\min_{\boldsymbol{G}_h} \frac{1}{(T-T_1-h+1)n} \left\|\boldsymbol{Y}_h^* - \hat{\boldsymbol{Y}}_h^* \boldsymbol{G}_h' \boldsymbol{S}'\right\|_F^2 + \lambda \|\,\text{vec}(\boldsymbol{G}_h)\|_1,$$

where $T_1$ denotes the minimum number of observations used for model training, $\|\cdot\|_F$ is the Frobenius norm, $\boldsymbol{Y}_h^* = [\boldsymbol{y}_{T_1+h}, \ldots, \boldsymbol{y}_T]' \in \mathbb{R}^{(T-T_1-h+1)\times n}$, $\hat{\boldsymbol{Y}}_h^* = \left[\hat{\boldsymbol{y}}_{T_1+h|T_1}, \ldots, \hat{\boldsymbol{y}}_{T|T-h}\right]' \in \mathbb{R}^{(T-T_1-h+1)\times n}$, and $\lambda \geq 0$ is a regularization parameter.

When $\lambda = 0$, the problem reduces to an empirical risk minimization (ERM) problem without regularization. Assuming that the series in the hierarchy are jointly weakly stationary and $\hat{\boldsymbol{Y}}_h^{*\prime}\hat{\boldsymbol{Y}}_h^*$ is invertible, it has a closed-form solution given by

$$\hat{\boldsymbol{G}}_h = \boldsymbol{B}_h^{*\prime}\hat{\boldsymbol{Y}}_h^* \left(\hat{\boldsymbol{Y}}_h^{*\prime}\hat{\boldsymbol{Y}}_h^*\right)^{-1},$$

where $\boldsymbol{B}_h^* = [\boldsymbol{b}_{T_1+h}, \dots, \boldsymbol{b}_T]' \in \mathbb{R}^{(T-T_1-h+1)\times n}$. If $\hat{\boldsymbol{Y}}_h^{*\prime}\hat{\boldsymbol{Y}}_h^*$ is not invertible, they suggested using a generalized inverse.

When $\lambda > 0$, imposing such a $L_1$ penalty on $\boldsymbol{G}_h$ will introduce sparsity and reduce estimation variance, albeit at the cost of introducing some bias. In addition, they also proposed another strategy that penalizes the matrix $\boldsymbol{G}_h$ towards the solution obtained by bottom-up method, i.e., $\boldsymbol{G}_{\text{BU}} = [\boldsymbol{0}_{n_b\times n_a} \mid \boldsymbol{I}_{n_b}]$.

Following the work, Wickramasuriya (2021) proposed an empirical MinT (EMinT) without the unbiasedness constraint by minimizing the trace of the covariance matrix of the reconciled forecast errors, $\text{Var}(\tilde{\boldsymbol{e}}_{T+h|T})$. Assuming that the series are jointly weakly stationary, she derived the solution given by

$$\hat{\boldsymbol{G}}_h = \boldsymbol{B}_h'\hat{\boldsymbol{Y}}_h \left(\hat{\boldsymbol{Y}}_h'\hat{\boldsymbol{Y}}_h\right)^{-1},$$

where $\boldsymbol{B}_h = [\boldsymbol{b}_h, \dots, \boldsymbol{b}_T]' \in \mathbb{R}^{(T-h+1)\times n}$, and $\hat{\boldsymbol{Y}}_h = \left[\hat{\boldsymbol{y}}_{h|0}, \dots, \hat{\boldsymbol{y}}_{T|T-h}\right]' \in \mathbb{R}^{(T-h+1)\times n}$. The difference between EMinT and ERM lies in the data sources used, as EMinT uses in-sample observations and base forecasts, while ERM relies on observations and base forecasts from a holdout validation set. We note that both ERM and EMinT consider an estimate of $\boldsymbol{G}$ that changes over the forecast horizon, which is why we keep the subscript $h$ here.

In practice, a prevalent challenge in forecast reconciliation arises when the base forecasts of some time series within the hierarchical structure may perform poorly, especially for large hierarchies. This can be attributed to either the inherent complexity of forecasting these series or potential model misspecification. In such cases, the effectiveness of forecast reconciliation may diminish, as the role of the weighting matrix $\boldsymbol{G}$ is to assimilate *all* base forecasts and map them into bottom-level disaggregated forecasts which are subsequently summed by $\boldsymbol{S}$. While the RERM method proposed by Ben Taieb & Koo (2019) introduces sparsity by shrinking some elements of $\boldsymbol{G}$ towards zero, it remains incapable of mitigating the adverse impact of underperforming base forecasts on the quality of the reconciled forecasts. Moreover, the method is time-consuming

because it uses expanding windows to recursively generate out-of-sample base forecasts, which are then used in the minimization problem.

We therefore propose two branches of innovative methods, constrained (out-of-sample-based) and unconstrained (in-sample-based) reconciliation with selection. These methods aim to identify and address the negative effect of some base forecasts of poor performance in a hierarchy on the overall performance of the reconciled forecasts. Additionally, through the incorporation of regularization in our objective function, our method has the potential to enhance reconciliation outcomes produced by using a "bad" choice of $\boldsymbol{W}$, thus reducing the risk of choosing estimator of $\boldsymbol{W}$. Moreover, our method generalizes to grouped hierarchies.

## 3 Forecast reconciliation with time series selection

In this section, we introduce our methods for keeping forecasts of an automatically selected set of series, identified as harmful to reconciliation, unused in forming reconciled forecasts, i.e., forecast reconciliation with series selection. Section 3.1 introduces constrained reconciliation methods with selection that formulate the problem based on out-of-sample base forecasts, while Section 3.2 presents an unconstrained reconciliation method with selection that formulates the problem based on in-sample observations and base forecasts.

### 3.1 Series selection with unbiasedness constraint

As $\boldsymbol{S}$ is fixed and $\hat{\boldsymbol{y}}_{T+h|T}$ is given, the estimation of $\boldsymbol{G}$ carries the linear reconciliation performance, as shown in Equation 1. (Subscript $h$ is dropped as we assume $\boldsymbol{W}$ and $\boldsymbol{G}$ do not change over the forecast horizon.) A natural way to keep forecasts of some series unused in reconciliation is through controlling the number of nonzero column entries in $\boldsymbol{G}$. This leads to a generalization of the MinT minimization problem by applying an additional penalty to the objective function. More precisely, we consider the optimization problem given by

$$\begin{aligned} \min_{\boldsymbol{G}} \quad & \frac{1}{2}\left(\hat{\boldsymbol{y}}_{T+h|T}-\boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{y}}_{T+h|T}\right)'\boldsymbol{W}^{-1}\left(\hat{\boldsymbol{y}}_{T+h|T}-\boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{y}}_{T+h|T}\right)+\lambda\mathfrak{g}(\boldsymbol{G}) \\ \text{s.t.} \quad & \boldsymbol{G}\boldsymbol{S}=\boldsymbol{I}, \end{aligned} \tag{3}$$

where $\mathfrak{g}(\cdot)$ is defined as an exterior penalty function designed to penalize the columns of $\boldsymbol{G}$ towards zero, with $\lambda$ is the corresponding penalty coefficient. Thus, this can be considered as a grouped variable selection problem, with each group corresponding to a column of $\boldsymbol{G}$. The constraint, $\boldsymbol{G}\boldsymbol{S}=\boldsymbol{I}$, reflects the assumption that the base forecasts and reconciled forecasts are

unbiased. When $\lambda = 0, \forall h$, the problem reduces to the MinT optimization problem in Equation 2 with a closed-form solution given by Equation 1.

**Proposition 1.** *Under the assumption of unbiasedness, the count of nonzero column entries of* $\boldsymbol{G}$ (*i.e., the number of time series selected for reconciliation*), *derived through solving Equation 3, is at least equal to the number of time series at the bottom level. In addition, we can restore the full hierarchical structure by aggregating/disaggregating the selected time series.*

*Proof.* According to the unbiasedness constraint $\boldsymbol{GS} = \boldsymbol{I}$, we have

$$\min\left(\operatorname{rank}(\boldsymbol{G}), \operatorname{rank}(\boldsymbol{S})\right) \geq \operatorname{rank}(\boldsymbol{I}_{n_b}) = n_b,$$

which indicates that the count of nonzero column entries of $\boldsymbol{G}$ is at least equal to $n_b$.

Let $\boldsymbol{X}_{\cdot\mathbb{S}} \in \mathbb{R}^{r\times|\mathbb{S}|}$ denote the submatrix of the $r \times c$ matrix $\boldsymbol{X}$ with column indices forming a set $\mathbb{S}$ (and when $\mathbb{S} = \{j\}$, we simply use $\boldsymbol{X}_{\cdot j}$). Here, $|\mathbb{S}|$ denotes the size of the set $\mathbb{S}$. Similarly, let $\boldsymbol{X}_{\mathbb{S}\cdot} \in \mathbb{R}^{|\mathbb{S}|\times c}$ denote the submatrix of $\boldsymbol{X}$ whose rows are indexed by a set $\mathbb{S}$ (and when $\mathbb{S} = \{i\}$, we simply use $\boldsymbol{X}_{i\cdot}$). Assuming that the set $\mathbb{S}$ involves the indices of nonzero columns in the solution of Equation 3, the following equations hold:

$$\begin{aligned}&\boldsymbol{GS} = \boldsymbol{G}_{\cdot\mathbb{S}}\boldsymbol{S}_{\mathbb{S}\cdot} \text{ and}\\ &\min\left(\operatorname{rank}(\boldsymbol{G}_{\cdot\mathbb{S}}), \operatorname{rank}(\boldsymbol{S}_{\mathbb{S}\cdot})\right) \geq \operatorname{rank}(\boldsymbol{I}_{n_b}) = n_b.\end{aligned}$$

Additionally, we have $\operatorname{rank}(\boldsymbol{S}_{\mathbb{S}\cdot}) \leq n_b$ as $\boldsymbol{S}$ has $n_b$ columns. Therefore, we can conclude that $\operatorname{rank}(\boldsymbol{S}_{\mathbb{S}\cdot}) = n_b$, which implies that the hierarchical structure can be fully restored by aggregating/disaggregating the selected time series, $(\boldsymbol{y}_t)_{\mathbb{S}}$.

For example, consider the simple hierarchy shown in Figure 1, it is not possible for our constrained reconciliation methods with selection to simultaneously zero out columns of $\boldsymbol{G}$ associated with series AA and AB. However, it is possible to zero out columns related to series AA and BA simultaneously.

**Proposition 2.** *The optimization problem in Equation 3 can be reformulated as a least squares problem with regularization and linear equality constraint as follows:*

$$\begin{aligned}\min_{\text{vec}(\boldsymbol{G})} \quad & \frac{1}{2}\left(\hat{\boldsymbol{y}}_{T+h|T}-\left(\hat{\boldsymbol{y}}'_{T+h|T}\otimes\boldsymbol{S}\right)\text{vec}(\boldsymbol{G})\right)'\boldsymbol{W}^{-1}\left(\hat{\boldsymbol{y}}_{T+h|T}-\left(\hat{\boldsymbol{y}}'_{T+h|T}\otimes\boldsymbol{S}\right)\text{vec}(\boldsymbol{G})\right)+\lambda\mathfrak{g}\left(\text{vec}(\boldsymbol{G})\right)\\ \text{s.t.} \quad & \left(\boldsymbol{S}'\otimes\boldsymbol{I}_{n_b}\right)\text{vec}(\boldsymbol{G})=\text{vec}(\boldsymbol{I}_{nb}),\end{aligned} \tag{4}$$

*which is characterized as a high-dimensional problem in which the number of features, denoted as $p = n_b \times n$, is much larger than the number of observations, $n$.*

*Proof.* Let $\text{vec}(\boldsymbol{A})$ denote the vectorization of a matrix $\boldsymbol{A}$, which stacks the columns of $\boldsymbol{A}$ on top of one another. We have

$$\begin{aligned}&\text{vec}\left(\hat{\boldsymbol{y}}_{T+h|T}\right)=\hat{\boldsymbol{y}}_{T+h|T},\\ &\text{vec}\left(\boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{y}}_{T+h|T}\right)=\left(\hat{\boldsymbol{y}}'_{T+h|T}\otimes\boldsymbol{S}\right)\text{vec}(\boldsymbol{G}),\\ &\text{vec}\left(\boldsymbol{G}\boldsymbol{S}\right)=\text{vec}\left(\boldsymbol{I}_{nb}\boldsymbol{G}\boldsymbol{S}\right)=\left(\boldsymbol{S}'\otimes\boldsymbol{I}_{n_b}\right)\text{vec}(\boldsymbol{G}).\end{aligned}$$

Substituting the terms in Equation 3 with these expressions, the previous problem now takes the form of a regression problem with an additional regularization term and an equality constraint on the coefficients, as shown in Equation 4.

Moving forward, we present three classes of regularizations we use to establish forecast reconciliation with series selection, resulting in the consideration of three optimization problems: (i) group best-subset selection with ridge regularization, (ii) intuitive method with $L_0$ regularization, and (iii) group lasso method.

### 3.1.1 Group best-subset selection with ridge regularization

In high-dimensional regime with $p \gg n$, a common desiderata is to assume that the true regression coefficient (i.e., $\text{vec}(\boldsymbol{G})$ in our problem) is sparse. We propose to apply a combination of $L_0$ and $L_2$ regularization as the exterior penalty function to control the nonzero column entries in $\boldsymbol{G}$:

$$\begin{aligned}\min_{\text{vec}(\boldsymbol{G})} \quad & \frac{1}{2}\left(\hat{\boldsymbol{y}}_{T+h|T}-\left(\hat{\boldsymbol{y}}'_{T+h|T}\otimes\boldsymbol{S}\right)\text{vec}(\boldsymbol{G})\right)'\boldsymbol{W}^{-1}\left(\hat{\boldsymbol{y}}_{T+h|T}-\left(\hat{\boldsymbol{y}}'_{T+h|T}\otimes\boldsymbol{S}\right)\text{vec}(\boldsymbol{G})\right)\\ & +\lambda_0\sum_{j=1}^{n}\mathbf{1}\left(\boldsymbol{G}_{\cdot j}\neq\boldsymbol{0}\right)+\lambda_2\left\|\text{vec}\left(\boldsymbol{G}\right)\right\|_2^2\\ \text{s.t.} \quad & \left(\boldsymbol{S}'\otimes\boldsymbol{I}_{n_b}\right)\text{vec}(\boldsymbol{G})=\text{vec}(\boldsymbol{I}_{nb}),\end{aligned} \tag{5}$$

where $\mathbf{1}(\cdot)$ is the indicator function, $\lambda_0 \geq 0$ controls the number of nonzero groups selected (each group corresponds to a column of $\boldsymbol{G}$), and $\lambda_2 \geq 0$ controls the strength of the ridge regularization. In a hierarchical time series context, the parameter of interest in Equation 5, $\text{vec}(\boldsymbol{G})$, has an inherent non-overlapping group structure, wherein each group corresponds to a single column of $\boldsymbol{G}$, each with a size of $n_b$. Therefore, we refer to this reconciliation method as *group best-subset selection with ridge regularization*. In the results that follow, we label the method differently based on various estimators for $\boldsymbol{W}$, referring to them as **OLS-subset**, **WLSs-subset**, **WLSv-subset**, **MinT-subset**, and **MinTs-subset**, respectively.

The inclusion of the ridge term in Equation 5 is motivated by earlier work on best-subset selection (Hazimeh & Mazumder 2020; Mazumder, Radchenko & Dedieu 2022), which suggests that additional ridge regularization can mitigate the poor predictive performance of best-subset selection in the low signal-to-noise ratio (SNR) regimes.

We present a Big-M based mixed integer programming (MIP) formulation for problem in Equation 5 given by

$$
\begin{aligned}
\min_{\text{vec}(\boldsymbol{G}),\boldsymbol{z},\check{\boldsymbol{e}},\boldsymbol{g}^+} \quad & \frac{1}{2}\check{\boldsymbol{e}}'\boldsymbol{W}^{-1}\check{\boldsymbol{e}} + \lambda_0 \sum_{j=1}^{n} z_j + \lambda_2 \boldsymbol{g}^{+\prime}\boldsymbol{g}^+ \\
\text{s.t.} \quad & \left(\boldsymbol{S}' \otimes \boldsymbol{I}_{n_b}\right) \text{vec}(\boldsymbol{G}) = \text{vec}\left(\boldsymbol{I}_{n_b}\right) \\
& \hat{\boldsymbol{y}}_{T+h|T} - \left(\hat{\boldsymbol{y}}'_{T+h|T} \otimes \boldsymbol{S}\right) \text{vec}(\boldsymbol{G}) = \check{\boldsymbol{e}} \\
& \sum_{i=1}^{n_b} g^+_{i+(j-1)n_b} \leqslant \mathcal{M} z_j, \quad j \in [n] \\
& \boldsymbol{g}^+ \geqslant \text{vec}(\boldsymbol{G}) \\
& \boldsymbol{g}^+ \geqslant -\text{vec}(\boldsymbol{G}) \\
& z_j \in \{0,1\}, \quad j \in [n],
\end{aligned}
\tag{6}
$$

where $\mathcal{M}$ is a Big-M parameter (a-priori specified) that is sufficiently large such that some optimal solution, say $\boldsymbol{g}^{+*}$, to Equation 6 satisfies $\max_{j\in[n]} \sum_{i=1}^{n_b} g^+_{i+(j-1)n_b} \leqslant \mathcal{M}$, the binary variable $z_j$ controls whether all the regression coefficients $\text{vec}(\boldsymbol{G})$ in group $j$ are zero or not, i.e., $z_j = 0$ implies that $\boldsymbol{G}_{\cdot j} = \boldsymbol{0}$, and $z_j = 1$ implies that $\sum_{i=1}^{n_b} g^+_{i+(j-1)n_b} \leqslant \mathcal{M}$. Such Big-M formulations are commonly used in MIP problems to model relations between discrete and continuous variables, and have been recently explored in regression with $L_0$ regularization. The problem is a mixed integer quadratic program (MIQP) that can be solved using commercial MIP solvers, e.g., Gurobi and CPLEX.

**Parameter tuning.** $\lambda_0 \geq 0$ and $\lambda_2 \geq 0$ are tuning parameters. To avoid computationally-expensive cross-validation, we tune the parameters to minimize the sum of squared reconciled forecast errors on the truncated training set, comprising only the $h$ observations closest to the forecast origin. Let $\lambda_0^1 = \frac{1}{2}\left(\hat{\boldsymbol{y}}_{T+h|T} - \tilde{\boldsymbol{y}}^{\text{bench}}_{T+h|T}\right)' \boldsymbol{W}^{-1}\left(\hat{\boldsymbol{y}}_{T+h|T} - \tilde{\boldsymbol{y}}^{\text{bench}}_{T+h|T}\right)$ that captures the scale of first term in the objective, where $\tilde{\boldsymbol{y}}^{\text{bench}}_{T+h|T}$ is a vector of reconciled forecasts obtained using Equation 1 with same estimator of $\boldsymbol{W}$, and define $\lambda_0^k = 0.0001\lambda_0^1$. For the parameter $\lambda_0$, we consider a grid of $k+1$ values, $\{\lambda_0^1, ..., \lambda_0^k, 0\}$, where $\lambda_0^j = \lambda_0^1 \left(\lambda_0^k/\lambda_0^1\right)^{(j-1)/(k-1)}$ for $j \in [k]$. So $\lambda_0^1, ..., \lambda_0^k$ is a sequence decreasing on the log scale. We use a grid of six values for the parameter $\lambda_2$, i.e., $\{0, 10^{-2}, 10^{-1}, 10^0, 10^1, 10^2\}$. Therefore, we tune over a two-dimensional grid of $(k+1) \times 6$ values to find the optimal combination of $\lambda_0$ and $\lambda_2$.

**Computation details.** The MIQP problem in Equation 6 is NP-Hard and computationally intensive. Bertsimas, King & Mazumder (2016) showed that commercial MIP solvers are capable of tackling problem instances for $p$ up to a thousand. To address larger instances, there has been impressive work on developing MIP-based approches for solving $L_0$-regularized regression problem, e.g., Bertsimas, King & Mazumder (2016), Hazimeh & Mazumder (2020), and Hazimeh, Mazumder & Saab (2022). However, it is challenging to extend their approaches to accommodate additional constraints within the optimization problem. Despite the potential sluggishness of handling large instances with commercial MIP solvers, in our experiments, we use Gurobi to solve our problem in Equation 6 by configuring parameters such as MIPGap = 0.001 and TimeLimit = 600 seconds for cases with $p > 1000$. So we can stop the solver before reaching the global optimum and obtain a suboptimal solution. This strategy is motivated by our need to consider numerous parameter candidates, with the final solution being validated against the training set, thus preventing the utilization of bad estimates of $\boldsymbol{G}$.

### 3.1.2 Intuitive method with $L_0$ regularization

Instead of estimating the entire matrix $\boldsymbol{G}$ in Section 3.1.1, we leverage the MinT solution in Equation 1 to streamline the optimization under consideration. Specifically, we define $\bar{\boldsymbol{S}} = \boldsymbol{A}\boldsymbol{S}$, where $\boldsymbol{A} = \text{diag}(\boldsymbol{z})$ is an $n \times n$ diagonal matrix, and $\boldsymbol{z}$ is an $n$-dimensional vector with elements either equal to 0 or 1. Taking the MinT solution in Equation 1, we have $\bar{\boldsymbol{G}} = (\boldsymbol{S}'\boldsymbol{A}'\boldsymbol{W}^{-1}\boldsymbol{A}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{A}'\boldsymbol{W}^{-1}$. Given fixed $\boldsymbol{S}$ and estimation of $\boldsymbol{W}$, $\bar{\boldsymbol{G}}$ is entirely determined by $\boldsymbol{A}$. By this way, when the $j$th diagnal element of $\boldsymbol{A}$ equals zero, the $j$th column of $\bar{\boldsymbol{G}}$ becomes entirely composed of zeros. Therefore, the optimization problem can be reduced to an integer quadratic programming (IQP) problem in which all of the variables are restricted to be integers:

$$\begin{aligned}
\min_{\boldsymbol{A}} \quad & \frac{1}{2}\left(\hat{\boldsymbol{y}}_{T+h|T}-\boldsymbol{S}\bar{\boldsymbol{G}}\hat{\boldsymbol{y}}_{T+h|T}\right)'\boldsymbol{W}^{-1}\left(\hat{\boldsymbol{y}}_{T+h|T}-\boldsymbol{S}\bar{\boldsymbol{G}}\hat{\boldsymbol{y}}_{T+h|T}\right)+\lambda_0\sum_{j=1}^{n}\boldsymbol{A}_{jj}\\
\text{s.t.} \quad & \bar{\boldsymbol{G}}=(\boldsymbol{S}'\boldsymbol{A}'\boldsymbol{W}^{-1}\boldsymbol{A}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{A}'\boldsymbol{W}^{-1}\\
& \bar{\boldsymbol{G}}\boldsymbol{S}=\boldsymbol{I},
\end{aligned}$$

where $\lambda_0 \geq 0$ controls the number of nonzero diagonal elements in $\boldsymbol{A}$, consequently affecting the number of nonzero columns (i.e., selected time series) in $\boldsymbol{G}$. We refer to this reconciliation method as *intuitive method with $L_0$ regularization*. In the results that follow, we label the method differently based on various estimators for $\boldsymbol{W}$, referring to them as **OLS-intuitive**, **WLSs-intuitive**, **WLSv-intuitive**, **MinT-intuitive**, and **MinTs-intuitive**, respectively.

We should note that implementing grouped variable selection with this optimization problem can be challenging because it imposes restrictions on the parameter of interest ($\bar{\boldsymbol{G}}$) to ensure it adheres rigorously to the analytical solution of MinT while making the selection.

To ensure the invertibility of $\boldsymbol{S}'\boldsymbol{A}'\boldsymbol{W}^{-1}\boldsymbol{A}\boldsymbol{S}$ and make the problem compatible with Gurobi, we reformulate the problem given by

$$\begin{aligned}
\min_{\boldsymbol{A},\bar{\boldsymbol{G}},\boldsymbol{C},\check{\boldsymbol{e}},\boldsymbol{z}} \quad & \frac{1}{2}\check{\boldsymbol{e}}'\boldsymbol{W}^{-1}\check{\boldsymbol{e}}+\lambda_0\sum_{j=1}^{n}z_j\\
\text{s.t.} \quad & \bar{\boldsymbol{G}}\boldsymbol{S}=\boldsymbol{I}\\
& \hat{\boldsymbol{y}}_{T+h|T}-\left(\hat{\boldsymbol{y}}'_{T+h|T}\otimes\boldsymbol{S}\right)\operatorname{vec}(\bar{\boldsymbol{G}})=\check{\boldsymbol{e}}\\
& \bar{\boldsymbol{G}}\boldsymbol{A}\boldsymbol{S}=\boldsymbol{I}\\
& \bar{\boldsymbol{G}}=\boldsymbol{C}\boldsymbol{S}'\boldsymbol{A}'\boldsymbol{W}^{-1}\\
& z_j\in\{0,1\}, \quad j\in[n].
\end{aligned} \tag{7}$$

**Parameter tuning.** Similarly to the setup in Section 3.1.1, we select the tuning parameter, $\lambda_0$, by minimizing the sum of squared reconciled forecast errors on a truncated training set, comprising only the $h$ observations occurred prior to the forecast origin. Let $\lambda_0^1 = \frac{1}{2}\left(\hat{\boldsymbol{y}}_{T+h|T}-\tilde{\boldsymbol{y}}^{\text{bench}}_{T+h|T}\right)'\boldsymbol{W}^{-1}\left(\hat{\boldsymbol{y}}_{T+h|T}-\tilde{\boldsymbol{y}}^{\text{bench}}_{T+h|T}\right)$, and $\lambda_0^k = 0.0001\lambda_0^1$, the collection of candidate values for $\lambda_0$ is $\{\lambda_0^1,...,\lambda_0^k,0\}$, where $\lambda_0^j = \lambda_0^1\left(\lambda_0^k/\lambda_0^1\right)^{(j-1)/(k-1)}$ for $j\in[k]$.

**Computation details.** Following a setup akin to that in Section 3.1.1, we employ Gurobi to solve Equation 7 by configuring parameters such as MIPGap = 0.001 and TimeLimit = 600 seconds for problems with $p > 1000$.

### 3.1.3 Group lasso method

Lasso are another popular methods for selection and estimation of parameters in the context of linear regression. Yuan & Lin (2006) introduced the group lasso method that can be used when there are grouped structure among the variables. Here, we consider *a group lasso problem under the unbiasedness assumption* given by

$$\begin{aligned} \min_{\boldsymbol{G}} \quad & \frac{1}{2}\left(\hat{\boldsymbol{y}}_{T+h|T}-\left(\hat{\boldsymbol{y}}'_{T+h|T}\otimes\boldsymbol{S}\right)\operatorname{vec}(\boldsymbol{G})\right)'\boldsymbol{W}^{-1}\left(\hat{\boldsymbol{y}}_{T+h|T}-\left(\hat{\boldsymbol{y}}'_{T+h|T}\otimes\boldsymbol{S}\right)\operatorname{vec}(\boldsymbol{G})\right) \\ & +\lambda\sum_{j=1}^{n}w_j\left\|\boldsymbol{G}_{\cdot j}\right\|_2 \\ \text{s.t.} \quad & \left(\boldsymbol{S}'\otimes\boldsymbol{I}_{n_b}\right)\operatorname{vec}(\boldsymbol{G})=\operatorname{vec}\left(\boldsymbol{I}_{n_b}\right), \end{aligned} \tag{8}$$

where $\lambda \geq 0$ is a tuning parameter, $w_j \neq 0$ is the penalty weight assigned in $\boldsymbol{G}_{\cdot j}$ to make model more flexible, and the second term in the objective is the penalty function that is intermediate between the $L_1$-penalty that is used in the lasso and the $L_2$-penalty that is used in ridge regression. In the results that follow, we label the method differently based on various estimators for $\boldsymbol{W}$, referring to them as **OLS-lasso**, **WLSs-lasso**, **WLSv-lasso**, **MinT-lasso**, and **MinTs-lasso**, respectively.

Next, we present the second order cone programming (SOCP) formulation for the group lasso based estimators given by

$$\begin{aligned} \min_{\operatorname{vec}(\boldsymbol{G}),\check{\boldsymbol{e}},\boldsymbol{g}^+} \quad & \frac{1}{2}\check{\boldsymbol{e}}'\boldsymbol{W}_h^{-1}\check{\boldsymbol{e}}+\lambda\sum_{j=1}^{n}w_jc_j \\ \text{s.t.} \quad & \left(\boldsymbol{S}'\otimes\boldsymbol{I}_{n_b}\right)\operatorname{vec}(\boldsymbol{G})=\operatorname{vec}\left(\boldsymbol{I}_{n_b}\right) \\ & \hat{\boldsymbol{y}}_{T+h|T}-\left(\hat{\boldsymbol{y}}'_{T+h|T}\otimes\boldsymbol{S}\right)\operatorname{vec}(\boldsymbol{G})=\check{\boldsymbol{e}} \\ & c_j=\sqrt{\sum_{i=1}^{n_b}g^2_{i+(j-1)n_b}}, \quad j\in[n]. \end{aligned} \tag{9}$$

Equation 9 includes additional auxiliary variables $c_j \in \mathbb{R}_{\geq 0}$, $j \in [n]$, and second order cone constraints, $c_j = \sqrt{\sum_{i=1}^{n_b} g^2_{i+(j-1)n_b}}$ for $j \in [n]$.

Compared to the previous two methods we proposed, the group lasso method is computationally friendlier. Nonetheless, Hazimeh, Mazumder & Radchenko (2023) demonstrated, both empirically and theoretically, that group $L_0$-regularized method exhibits advantages over its group lasso counterpart across a range of regimes. Group lasso can either be highly dense or possess non-zero coefficients that are overly shrunk. This issue becomes more pronounced when

the groups are correlated with each other as group lasso tends to retain all correlated groups instead of seeking a more concise model.

**Penalty weights and parameter tuning.** In the context of group lasso, the default choice for the penalty weight, $w_j$, is $\sqrt{p_j}$, where $p_j$ is the size of each group (in our case, $p_j = n_b$). In our experiment, we allocate different penalty weights to each group by considering $w_j = 1/\left\|\boldsymbol{G}_{\cdot j}^{\text{bench}}\right\|_2$, which allows us to account for variations in scale across different levels in the hierarchy.

We compute the group lasso over $k+1$ values of the tuning parameter $\lambda$, and select the tuning parameter by optimizing the sum of squared reconciled forecast errors on a truncated training set, consisting only of $h$ observations occurred prior to the forecast origin. The collection of candidate values for $\lambda$ under consideration is $\{\lambda^1, ..., \lambda^k, 0\}$, where $\lambda^1 = \max_{j=1,\dots,n} \left\| -\left(\left(\hat{\boldsymbol{y}}'_{T+h|T} \otimes \boldsymbol{S}\right)_{\cdot j^*}\right)' \boldsymbol{W}^{-1} \hat{\boldsymbol{y}}_{T+h|T} \right\|_2 / w_j$, $\lambda^k = 0.0001\lambda^1$, and $\lambda^j = \lambda^1 \left(\lambda^k/\lambda^1\right)^{(j-1)/(k-1)}$ for $j \in [k]$.

**Proposition 3.** *Ignoring the unbiasedness constraint, we define $\lambda^1$ as the smallest $\lambda$ value such that all predictors in the group lasso problem have zero coefficients. Then we have*

$$\lambda^1 = \max_{j=1,\dots,n} \left\| -\left(\left(\hat{\boldsymbol{y}}'_{T+h|T} \otimes \boldsymbol{S}\right)_{\cdot j^*}\right)' \boldsymbol{W}^{-1} \hat{\boldsymbol{y}}_{T+h|T} \right\|_2 / w_j,$$

*where $j^*$ denotes the column index of $\hat{\boldsymbol{y}}'_{T+h|T} \otimes \boldsymbol{S}$ that corresponds to the $j$th column of $\boldsymbol{G}$.*

*Proof.* Denote $\boldsymbol{\beta} = \text{vec}(\boldsymbol{G})$, and the first term in the objective of Equation 8 as $L\left(\boldsymbol{\beta} \mid \boldsymbol{D}\right)$, where $\boldsymbol{D}$ is the working data $\{\hat{\boldsymbol{y}}_{T+h|T}, \hat{\boldsymbol{y}}'_{T+h|T} \otimes \boldsymbol{S}\}$. Ignoring the unbiasedness constraint, we define $\lambda^1$ as the smallest $\lambda$ value such that all predictors in the group lasso problem have zero coefficients, i.e., the solution at $\lambda^1$ is $\hat{\boldsymbol{\beta}}^1 = \boldsymbol{0}$. (Note that there is no intercept in our problem.) Under the Karush-Kuhn-Tucker conditions, we have

$$\begin{aligned}
\lambda^1 &= \max_{j=1,\dots,n} \left\| \left[\nabla L\left(\hat{\boldsymbol{\beta}}^1 \mid \mathbf{D}\right)\right]^{(j)} \right\|_2 / w_j \\
&= \max_{j=1,\dots,n} \left\| -\left(\left(\hat{\boldsymbol{y}}'_{T+h|T} \otimes \boldsymbol{S}\right)_{\cdot j^*}\right)' \boldsymbol{W}^{-1} \hat{\boldsymbol{y}}_{T+h|T} \right\|_2 / w_j.
\end{aligned}$$

**Computation details.** Due to the incorporation of the unbiasedness constraint, we can not directly use some open-source packages designed for group lasso. Consequently, we employ Gurobi to solve the SOCP problem in Equation 9, configuring it with setting OptimalityTol = 0.0001.

### 3.2 Series selection method without unbiasedness constraint

In this section, we relax the unbiasedness constraint, $\boldsymbol{GS} = \boldsymbol{I}$, and introduce a reconciliation method with selection that relies on in-sample observations and fitted values. Let $\boldsymbol{Y} \in \mathbb{R}^{T\times n}$ denote a matrix comprising observations from all time series on the training set in the structure, and $\hat{\boldsymbol{Y}} \in \mathbb{R}^{T\times n}$ denote a matrix of in-sample one-step-ahead forecasts (i.e., fitted values) for all time series, where $T$ is the length of the training set. The proposed *empirical group lasso* method considers the optimization problem given by

$$\min_{\boldsymbol{G}} \quad \frac{1}{2T}\left\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\boldsymbol{G}'\boldsymbol{S}'\right\|_F^2 + \lambda \sum_{j=1}^{n} w_j \left\|\boldsymbol{G}_{\cdot j}\right\|_2,$$

where $\|\cdot\|_F$ is the Frobenius norm, and $\lambda \geq 0$ is a tuning parameter, $w_j \neq 0$ is the penalty weight assigned in $\boldsymbol{G}_{\cdot j}$ to make a more flexible model. Following the work by Ben Taieb & Koo (2019), using the fact that $\|\boldsymbol{X}\|_F^2 = \|\operatorname{vec}(X)\|_2^2$ and the useful formulation that $\operatorname{vec}(\boldsymbol{ABC}) = (\boldsymbol{C}' \otimes \boldsymbol{A})\operatorname{vec}(\boldsymbol{B})$, we rewrite the problem as

$$\min_{\operatorname{vec}(\boldsymbol{G})} \quad \frac{1}{2N}\left\|\operatorname{vec}(\boldsymbol{Y}) - (\boldsymbol{S}\otimes\hat{\boldsymbol{Y}})\operatorname{vec}\left(\boldsymbol{G}'\right)\right\|_2^2 + \lambda \sum_{j=1}^{n} w_j \left\|\boldsymbol{G}_{\cdot j}\right\|_2,$$

which transforms into a standard group lasso problem, with $\operatorname{vec}(\boldsymbol{Y})$ serving as the dependent variable and $\boldsymbol{S}\otimes\hat{\boldsymbol{Y}}$ as the covariate matrix. We denote this as **Elasso** in the results that follow.

We also explored the empirical version of group best-subset selection with ridge regularization and Intuitive method with $L_0$ regularization in which we do not impose the unbiasedness constraint. It is worth mentioning that Hazimeh, Mazumder & Radchenko (2023) presented a new algorithmic framework for formulating the group $L_0$ problem with ridge regularization and provided an open-source implementation of the algorithm available on github at https://github.com/hazimehh/L0Group. However, our experiments showed that this algorithm can not terminate within five hours for typical instances with $p \sim 10^4$. Therefore, in this paper, we only present the empirical group lasso method for series selection method without unbiasedness constraint.

**Penalty weights and parameter tuning.** Similarly to the setup in Section 3.1.3, we assign different penalty weights to each group by setting $w_j = 1/\left\|\boldsymbol{G}_{\cdot j}^{\mathrm{OLS}}\right\|_2$, where $\boldsymbol{G}^{\mathrm{OLS}}$ is the solution obtained by the OLS estimator of $\boldsymbol{W}$. We select the tuning parameter, $\lambda$, by minimizing the sum of squared reconciled forecast errors on a truncated training set, comprising only the $h$ observations closest to the forecast origin. Specifically, we form the set of candidate values for $\lambda$

as $\{\lambda^1, ..., \lambda^k, 0\}$, where $\lambda^1 = \max_{j=1,...,n} \left\| -\frac{1}{N} \left( \left( \boldsymbol{S} \otimes \hat{\boldsymbol{Y}} \right)_{\cdot j*} \right)' \text{vec}(\boldsymbol{Y}) \right\|_2 / w_j$ is the smallest $\lambda$ value such that all predictors in the empirical group lasso problem have zero coefficients, i.e., $\boldsymbol{G} = \boldsymbol{O}_{\cdot}$, $\lambda^k = 0.0001\lambda^1$, and $\lambda^j = \lambda^1 \left(\lambda^k/\lambda^1\right)^{(j-1)/(k-1)}$ for $j \in [k]$.

**Computation details.** While there are open-source packages available for solving a group lasso problem, they are still relatively slow when applied to large instances for practical usage. For example, given a specific value for tuning parameter, our experiments observed that we can not obtain a solution within two hours for typical instances with $p \sim 10^4$ when using the gglasso R package. Instead, we use Gurobi to solve the problem based on the SOCP formulation for the empirical group lasso. The formulation aligns with Equation 9 but omits the unbiasedness constraint.

## 4 Monte Carlo simulations

### 4.1 Model misspecification in a hierarchy

### 4.2 Exploring the effect of correlation

## 5 Applications

### 5.1 Forecasting Australian domestic tourism

### 5.2 Forecasting Australian labour force

## 6 Conclusion

### Acknowledgement

## References

Athanasopoulos, G, RJ Hyndman, N Kourentzes & F Petropoulos (2017). Forecasting with temporal hierarchies. *European Journal of Operational Research* **262**(1), 60–74.

Ben Taieb, S & B Koo (2019). Regularized Regression for Hierarchical Forecasting Without Unbiasedness Conditions. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '19. Anchorage, AK, USA: Association for Computing Machinery, pp.1337–1347.

Bertsimas, D, A King & R Mazumder (2016). Best subset selection via a modern optimization lens. en. *The Annals of Statistics* **44**(2), 813–852.

Hazimeh, H & R Mazumder (2020). Fast Best Subset Selection: Coordinate Descent and Local Combinatorial Optimization Algorithms. *Operations Research* **68**(5), 1517–1537.

Hazimeh, H, R Mazumder & P Radchenko (2023). Grouped variable selection with discrete optimization: Computational and statistical perspectives. en. *The Annals of Statistics* **51**(1), 1–32.

Hazimeh, H, R Mazumder & A Saab (2022). Sparse regression at scale: branch-and-bound rooted in first-order optimization. *Mathematical Programming* **196**(1), 347–388.

Hyndman, RJ, RA Ahmed, G Athanasopoulos & HL Shang (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis* **55**(9), 2579–2589.

Hyndman, RJ, AJ Lee & E Wang (2016). Fast computation of reconciled forecasts for hierarchical and grouped time series. *Computational Statistics & Data Analysis* **97**, 16–32.

Mazumder, R, P Radchenko & A Dedieu (2022). Subset Selection with Shrinkage: Sparse Linear Modeling When the SNR Is Low. *Operations Research*.

Panagiotelis, A, G Athanasopoulos, P Gamakumara & RJ Hyndman (2021). Forecast reconciliation: A geometric view with new insights on bias correction. *International Journal of Forecasting* **37**(1), 343–359.

Pritularga, KF, I Svetunkov & N Kourentzes (2021). Stochastic coherency in forecast reconciliation. *International Journal of Production Economics* **240**, 108221.

Wickramasuriya, SL (2021). Properties of point forecast reconciliation approaches. arXiv: 2103.11129 [stat.ME].

Wickramasuriya, SL, G Athanasopoulos & RJ Hyndman (2019). Optimal Forecast Reconciliation for Hierarchical and Grouped Time Series Through Trace Minimization. *Journal of the American Statistical Association* **114**(526), 804–819.

Yuan, M & Y Lin (2006). Model selection and estimation in regression with grouped variables. en. *Journal of the Royal Statistical Society. Series B, Statistical Methodology* **68**(1), 49–67.