

Hierarchical Forecasts Reconciliation

Puwasala Gamakumara*

Department of Econometrics and Business Statistics,
Monash University,
VIC 3800, Australia.

Email: Puwasala.Gamakumara@monash.edu

and

Anastasios Panagiotelis

Department of Econometrics and Business Statistics,
Monash University,
VIC 3800, Australia.

Email: Anastasios.Panagiotelis@monash.edu

and

George Athanasopoulos

Department of Econometrics and Business Statistics,
Monash University,
VIC 3800, Australia.

Email: george.athanasopoulos@monash.edu

and

Rob J Hyndman

Department of Econometrics and Business Statistics,
Monash University,
VIC 3800, Australia.

Email: rob.hyndman@monash.edu

June 28, 2019

Abstract

TBC

*The authors gratefully acknowledge the support of Australian Research Council Grant DP140103220. We also thank Professor Mervyn Silvapulle for valuable comments.

1 Introduction

2 Coherent forecasts

2.1 Notation and preliminaries

We briefly define the concept of a *hierarchical time series* in a fashion similar to [, before](#) citations elaborating on some of the limitations of this understanding. A *hierarchical time series* is a collection of n variables indexed by time, where some variables are aggregates of other variables. We let $\mathbf{y}_t \in \mathbb{R}^n$ be a vector comprising observations of all variables in the hierarchy at time t . The *bottom-level series* are defined as those m variables that cannot be formed as aggregates of other variables; we let $\mathbf{b}_t \in \mathbb{R}^m$ be a vector comprised of observations of all bottom-level series at time t . The hierarchical structure of the data implies that

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

where \mathbf{S} is an $n \times m$ constant matrix that encodes the aggregation constraints, holds for all t .

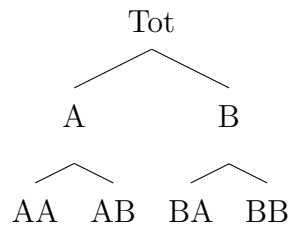


Figure 1: An example of a two level hierarchical structure.

To clarify these concepts consider the example of the hierarchy in Figure 1. For this hierarchy, $n = 7$, $\mathbf{y}_t = [y_{Tot,t}, y_{A,t}, y_{B,t}, y_{C,t}, y_{AA,t}, y_{AB,t}, y_{BA,t}, y_{BB,t}]'$, $m = 4$, $\mathbf{b}_t = [y_{AA,t}, y_{AB,t}, y_{BA,t}, y_{BB,t}]'$ and

$$\mathbf{S} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ \mathbf{I}_4 \end{pmatrix},$$

where \mathbf{I}_4 is the 4×4 identity matrix.

While such a definition is completely serviceable, it obscures the full generality of the methodologies developed in the literature on hierarchical time series. There are only two important characteristics of the data that we are interested in; the first is that they are multivariate, the second is that they adhere to linear constraints. Below we provide definitions that provide geometric intuition behind the data and methodologies employed in the hierarchical forecasting.

2.2 Coherence

Definition 2.1 (Coherent subspace). The m -dimensional linear subspace $\mathfrak{s} \subset \mathbb{R}^n$ for which a set of linear constraints hold for all $\mathbf{y} \in \mathfrak{s}$ is defined as the *coherent subspace*.

To further illustrate, Figure 2 depicts the most simple three variable hierarchy where $y_{Tot,t} = y_{A,t} + y_{B,t}$. The coherent subspace is depicted as a grey 2-dimensional plane within 3-dimensional space, i.e. $m = 2$ and $n = 3$. It is worth noting that the coherent subspace is spanned by the columns of \mathbf{S} , i.e. $\mathfrak{s} = \text{span}(\mathbf{S})$. In Figure 2, these columns are $\vec{s}_1 = (1, 1, 0)'$ and $\vec{s}_2 = (1, 0, 1)'$. However, it is equally important to recognise that the hierarchy could also have been defined in terms of $y_{Tot,t}$ and $y_{A,t}$ rather than the bottom level series, $y_{A,t}$ and

$y_{B,t}$. In this case the corresponding ‘ \mathbf{S} matrix’ would have columns $(1, 0, 1)'$ and $(0, 1, -1)'$, which also span the coherent subspace. Thus, while there are multiple ways to define a hierarchy and an associated ‘ \mathbf{S} matrix’, the columns of any ‘ \mathbf{S} matrix’ will always span the same unique coherent subspace.

Also notable by its absence in the above definition is any reference to *aggregation*. As the literature has shown, the linear constraints need not be aggregation constraints at all. For example consider weighted sums, while consider an example where one variable is the difference of two other variables.

find
refer-
ence

include
Li and
Tang
refer-
ence

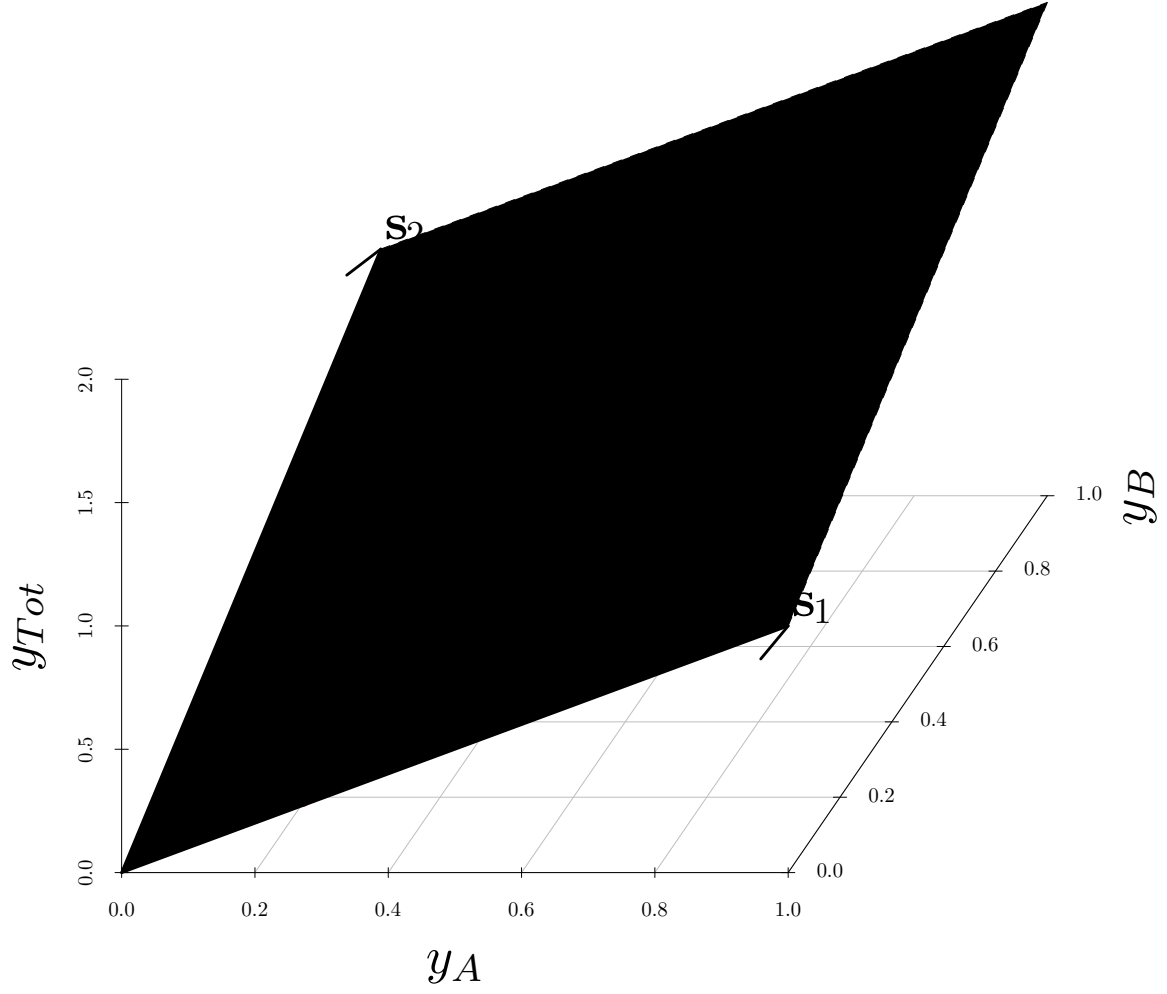


Figure 2: Depiction of a three dimensional hierarchy with $y_{\text{Tot}} = y_A + y_B$. The gray colour two dimensional plane reflects the coherent subspace \mathfrak{s} where $\vec{s}_1 = (1, 1, 0)'$ and $\vec{s}_2 = (1, 0, 1)'$ are basis vectors that spans \mathfrak{s} . The points in \mathfrak{s} represents realisations or coherent forecasts

Definition 2.2 (Hierarchical Time Series). A hierarchical time series is an n -dimensional multivariate time series such that all observed values $\mathbf{y}_1, \dots, \mathbf{y}_T$ and all future values $\mathbf{y}_{T+1}, \mathbf{y}_{T+2}, \dots$ lie in the coherent subspace, i.e. $\mathbf{y}_t \in \mathfrak{s} \quad \forall t$.

Despite the common use of the term *hierarchical time series*, it should be clear from the definition that the data need not necessarily follow a hierarchy. In fact, the term *hierarchical* is misleading since the literature has covered instances that cannot easily be depicted as hierarchies as in Figure 1. These include . Furthermore, although the definition makes clear reference to time series, this definition can be easily generalised to any vector-valued data for which some constraints are known to hold.

include
refer-
ences

Definition 2.3 (Coherent Point Forecasts). Let $\check{\mathbf{y}}_{t+h|t} \in \mathbb{R}^n$ be a point forecast of the values of all series in the hierarchy at time $t+h$, made using information up to and including time t . Then $\check{\mathbf{y}}_{t+h|t}$ is *coherent* if $\check{\mathbf{y}}_{t+h|t} \in \mathfrak{s}$.

Without any loss of generality, that above definition could also be applied to prediction for multivariate data in general, rather than just forecasting of time series. While the observed data will be coherent by definition, it is important to note that there are a number of reasons why forecasts or predictions may be incoherent. First, since applications of hierarchical forecasting tend to be very high dimensional a common strategy in practice is to produce forecasts for each time series independently using univariate models. Second, even where a multivariate model is used for the full vector of observations, it may be difficult to capture the linear constraints inherent in the data particularly for complicated non-linear models. Third, in some cases judgemental adjustments may be made inducing incoherent forecasts.

3 Forecasts reconciliation

3.1 Point forecast reconciliation

As discussed, reconciliation is distinct from coherence, since the former refers to a process whereby incoherent forecasts are made coherent. Although reconciliation methods for point forecasts are extant in the literature they are rarely defined explicitly. We do so here in slightly more general terms than usual.

Let $\hat{\mathbf{y}}_{t+h|t} \in \mathbb{R}^n$ be any set of incoherent point forecasts at time $t + h$ conditional on information up to and including time t . We now introduce a linear function that converts unreconciled forecasts into new bottom level forecasts. Let \mathbf{G} and \mathbf{d} be an $m \times n$ matrix and $m \times 1$ vector respectively, and let $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be the mapping $g(\mathbf{y}) = \mathbf{G}\mathbf{y} + \mathbf{d}$. A composition of g and $s(\cdot)$ gives the following definition for point forecast reconciliation.

Definition 3.1. The point forecast $\tilde{\mathbf{y}}_{t+h|t}$ “reconciles” $\hat{\mathbf{y}}_{t+h|t}$ with respect to the mapping $g(\cdot)$ iff

$$\tilde{\mathbf{y}}_{t+h|t} = \mathbf{S}(\mathbf{G}\hat{\mathbf{y}}_{t+h|t} + \mathbf{d}). \quad (2)$$

Several choices of $g(\cdot)$ currently extant in the literature, including the OLS, WLS and MinT methods, are special cases where $s \circ g$ is a projection. These can be defined so that $\mathbf{G} = (\mathbf{R}'_{\perp} \mathbf{S})^{-1} \mathbf{R}'_{\perp}$ and $\mathbf{d} = \mathbf{0}$, where, \mathbf{R}_{\perp} is a $n \times m$ orthogonal complement to an $n \times (n - m)$ matrix \mathbf{R} , where the columns of the latter span the null space of \mathbf{S} . For example, a straightforward choice of \mathbf{R} for the most simple three variable hierarchy where $y_{1,t} = y_{2,t} + y_{3,t}$, is the vector $(1, -1, -1)$ which is orthogonal (in the Euclidean sense) to the columns of \mathbf{S} . In this case, the matrix \mathbf{R} can be interpreted as a ‘restrictions’ matrix since it has the property that $\mathbf{R}'\mathbf{y} = \mathbf{0}$ for coherent \mathbf{y} . In OLS reconciliation, $\mathbf{R}'_{\perp} = \mathbf{S}'$ whereas in MinT or WLS reconciliation \mathbf{R}'_{\perp} takes the form $\mathbf{S}'\mathbf{W}^{-1}$. We will be discussing

these projections distinctly in the next subsection.

The columns of \mathbf{S} and \mathbf{R} provide a basis for \mathbb{R}^n . Therefore any incoherent set of point forecasts $\hat{\mathbf{y}}_{t+h|t} \in \mathbb{R}^n$ can be expressed in terms of coordinates in the basis defined by \mathbf{S} and \mathbf{R} . Let $\tilde{\mathbf{b}}_{t+h|t}$ and $\tilde{\mathbf{a}}_{t+h|t}$ be the coordinates corresponding to \mathbf{S} and \mathbf{R} respectively, after a change of basis. The process of reconciliation involves setting the values of the reconciled bottom-level series to be $\tilde{\mathbf{b}}_{t+h|t}$, and ignoring $\tilde{\mathbf{a}}_{t+h|t}$ to ensure coherence. From properties of linear algebra it follows that

$$\hat{\mathbf{y}}_{t+h|t} = (\mathbf{S} \ \mathbf{R}) \begin{pmatrix} \tilde{\mathbf{b}}_{t+h|t} \\ \tilde{\mathbf{a}}_{t+h|t} \end{pmatrix} = \mathbf{S}\tilde{\mathbf{b}}_{t+h|t} + \mathbf{R}\tilde{\mathbf{a}}_{t+h|t},$$

while the reconciled point forecast is

$$\tilde{\mathbf{y}}_{t+h|t} = \mathbf{S}\tilde{\mathbf{b}}_{t+h|t}.$$

In order to find $\tilde{\mathbf{b}}_{t+h|t}$ we require the inverse $(\mathbf{S} \ \mathbf{R})^{-1}$ which is given by

$$(\mathbf{S} \ \mathbf{R})^{-1} = \begin{pmatrix} (\mathbf{R}'_{\perp} \mathbf{S})^{-1} \mathbf{R}'_{\perp} \\ (\mathbf{S}'_{\perp} \mathbf{R})^{-1} \mathbf{S}'_{\perp} \end{pmatrix}, \quad (3)$$

where \mathbf{S}_{\perp} is the orthogonal complements of \mathbf{S} . Thus it follows that $\tilde{\mathbf{b}}_{t+h|t} = (\mathbf{R}'_{\perp} \mathbf{S})^{-1} \mathbf{R}'_{\perp} \hat{\mathbf{y}}_{t+h|t}$ and $\tilde{\mathbf{y}}_{t+h|t} = \mathbf{S}(\mathbf{R}'_{\perp} \mathbf{S})^{-1} \mathbf{R}'_{\perp} \hat{\mathbf{y}}_{t+h|t}$. Here $(\mathbf{R}'_{\perp} \mathbf{S})^{-1} \mathbf{R}'_{\perp}$ corresponds to \mathbf{G} as defined previously.

3.2 Motivation of using projections

We have seen from the above discussion that projection is playing an important role in point forecast reconciliation. Now we turn our attention to the following two theorems that explains the motivation of using projection in this context.

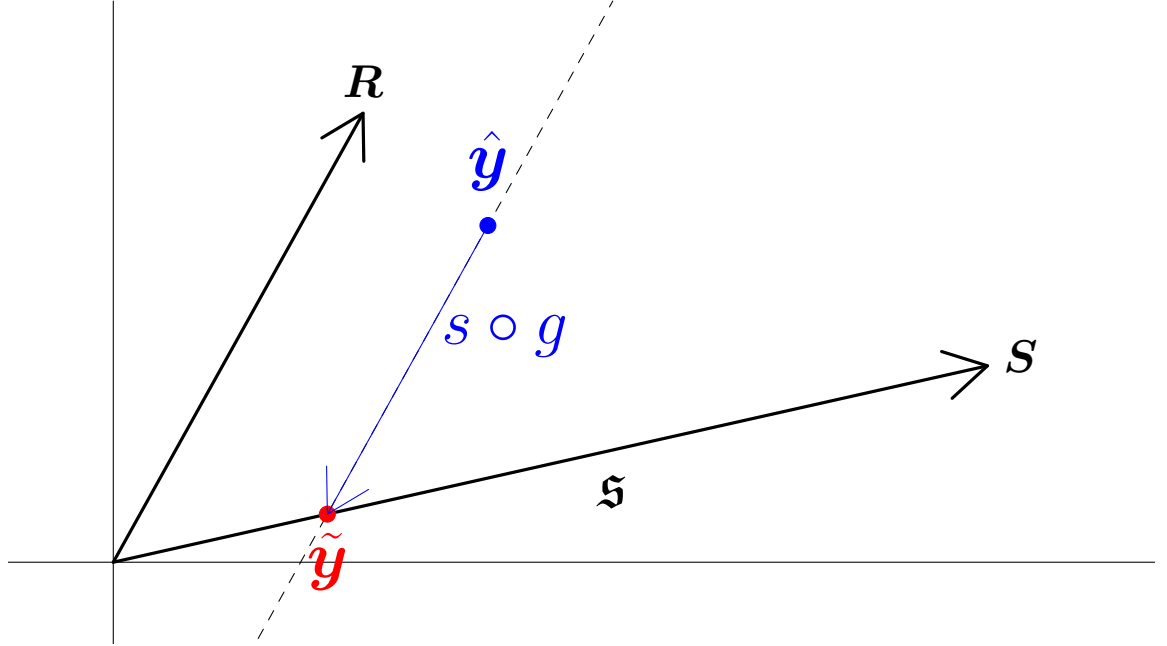


Figure 3: Summary of probabilistic point reconciliation. The mapping $s \circ g$ projects the unreconciled forecast $\hat{\mathbf{y}}_{t+h|h}$ onto \mathfrak{s} , yielding the reconciled forecast $\tilde{\mathbf{y}}_{t+h|h}$ with subscripts dropped in the figure for ease of presentation. Since the smallest hierarchy involves three dimensions, this figure is only a schematic.

First, let $\boldsymbol{\mu}_{t+h|t} := \mathbb{E}(\mathbf{y}_{t+h} \mid \mathbf{y}_1, \dots, \mathbf{y}_t)$ and assume $\hat{\mathbf{y}}_{t+h|t}$ is an unbiased prediction; that is $\mathbb{E}_{1:t}(\hat{\mathbf{y}}_{t+h|t}) = \boldsymbol{\mu}_{t+h|t}$, where the subscript $1:t$ denotes an expectation taken over the training sample.

Theorem 3.1 (Unbiasedness preserving property). *For unbiased $\hat{\mathbf{y}}_{t+h|t}$, the reconciled point forecast is also an unbiased prediction as long as $s \circ g$ is a projection onto \mathfrak{s} .*

Proof. The expected value of the reconciled forecast is given by

$$\mathbb{E}_{1:t}(\tilde{\mathbf{y}}_{t+h|t}) = \mathbb{E}_{1:t}(\mathbf{S}\mathbf{G}\hat{\mathbf{y}}_{t+h|t}) = \mathbf{S}\mathbf{G}\mathbb{E}_{1:t}(\hat{\mathbf{y}}_{t+h|t}) = \mathbf{S}\mathbf{G}\boldsymbol{\mu}_{t+h|t}.$$

Since the aggregation constraints hold for the true data generating process, $\boldsymbol{\mu}_{t+h|t}$ must lie in \mathfrak{s} . If $\mathbf{S}\mathbf{G}$ is a projection, then it is equivalent to the identity map for all vectors that lie in its range. Therefore $\mathbf{S}\mathbf{G}\boldsymbol{\mu}_{t+h|t} = \boldsymbol{\mu}_{t+h|t}$ when $\mathbf{S}\mathbf{G}$ is a projection matrix. \square

We note the above result holds when the projection $s \circ g$ is only onto the coherent subspace \mathfrak{s} . That is the result does not hold for any general g even when the range of $s \circ g$ is \mathfrak{s} . To describe this more explicitly suppose $s \circ g$ is a projection to any linear subspace \mathfrak{L} of \mathfrak{s} . Then $\mathbf{S}\mathbf{G}\boldsymbol{\mu}_{t+h|t} \neq \boldsymbol{\mu}_{t+h|t}$ as the projection will move $\boldsymbol{\mu}_{t+h|t}$ to a point $\bar{\boldsymbol{\mu}}$ in \mathfrak{L} as depicted in the Figure 4. Thus $\mathbb{E}_{1:t}(\tilde{\mathbf{y}}_{t+h|t}) \neq \boldsymbol{\mu}_{t+h|t}$ which breaks the unbiasedness. Recall the top-down method (Gross & Sohl 1990) with

$$\mathbf{G} = \begin{pmatrix} \mathbf{p} & \mathbf{0}_{(m \times n-1)} \end{pmatrix} \quad (4)$$

where $\mathbf{p} = (p_1, \dots, p_m)'$ is an m -dimensional vector consisting a set of proportions which is use to disaggregate the top-level forecasts along the hierarchy. Hyndman et al. (2011) claimed that this method is not producing unbiased coherent forecasts even if the base forecasts are unbiased since $\mathbf{S}\mathbf{G}\mathbf{S} \neq \mathbf{S}$ for \mathbf{G} in (4). However the more rational explanation is that, in top-down approach the projection $s \circ g$ is not onto \mathfrak{s} , but to a linear subspace of

\mathfrak{s} spanned by \mathbf{p} . Thus from above explanation it follows that $\mathbf{SG}\boldsymbol{\mu}_{t+h|t} \neq \boldsymbol{\mu}_{t+h|t}$ and hence not producing unbiased forecasts.

Now let \mathbf{y}_{t+h} be the realisation of the data generating process at time $t + h$, and let $\|\mathbf{v}\|_2$ be the L_2 norm of vector \mathbf{v} . The following theorem shows that reconciliation never increases, and in most cases reduces, the sum of squared errors of point forecasts.

Theorem 3.2 (Distance reducing property). *If $\tilde{\mathbf{y}}_{t+h|t} = \mathbf{SG}\hat{\mathbf{y}}_{t+h|t}$, where \mathbf{G} is such that \mathbf{SG} is an orthogonal projection onto \mathfrak{s} , then the following inequality holds:*

$$\|(\tilde{\mathbf{y}}_{t+h|t} - \mathbf{y}_{t+h})\|_2^2 \leq \|(\hat{\mathbf{y}}_{t+h|t} - \mathbf{y}_{t+h})\|_2^2. \quad (5)$$

Proof. Since the aggregation constraints must hold for all realisations, $\mathbf{y}_{t+h} \in \mathfrak{s}$ and $\mathbf{y}_{t+h} = \mathbf{SG}\mathbf{y}_{t+h}$ whenever \mathbf{SG} is a projection onto \mathfrak{s} . Therefore,

$$\|(\tilde{\mathbf{y}}_{t+h|t} - \mathbf{y}_{t+h})\|_2 = \|(\mathbf{SG}\hat{\mathbf{y}}_{t+h|t} - \mathbf{SG}\mathbf{y}_{t+h})\|_2 \quad (6)$$

$$= \|\mathbf{SG}(\hat{\mathbf{y}}_{t+h|t} - \mathbf{y}_{t+h})\|_2. \quad (7)$$

The Cauchy-Schwarz inequality can be used to show that orthogonal projections are bounded operators (Hunter & Nachtergaele 2001), therefore

$$\|\mathbf{SG}(\hat{\mathbf{y}}_{t+h|t} - \mathbf{y}_{t+h})\|_2 \leq \|(\hat{\mathbf{y}}_{t+h|t} - \mathbf{y}_{t+h})\|_2.$$

□

The inequality is strict whenever $\hat{\mathbf{y}}_{t+h|t} \notin \mathfrak{s}$.

Point reconciliation methods based on projections will always minimise the distance between unreconciled and reconciled forecasts, however the specific distance will depend on the choice of \mathbf{R} . Following subsections will explicitly discuss the different projection based reconciliation methods and their optimality based on distinct distance measures.

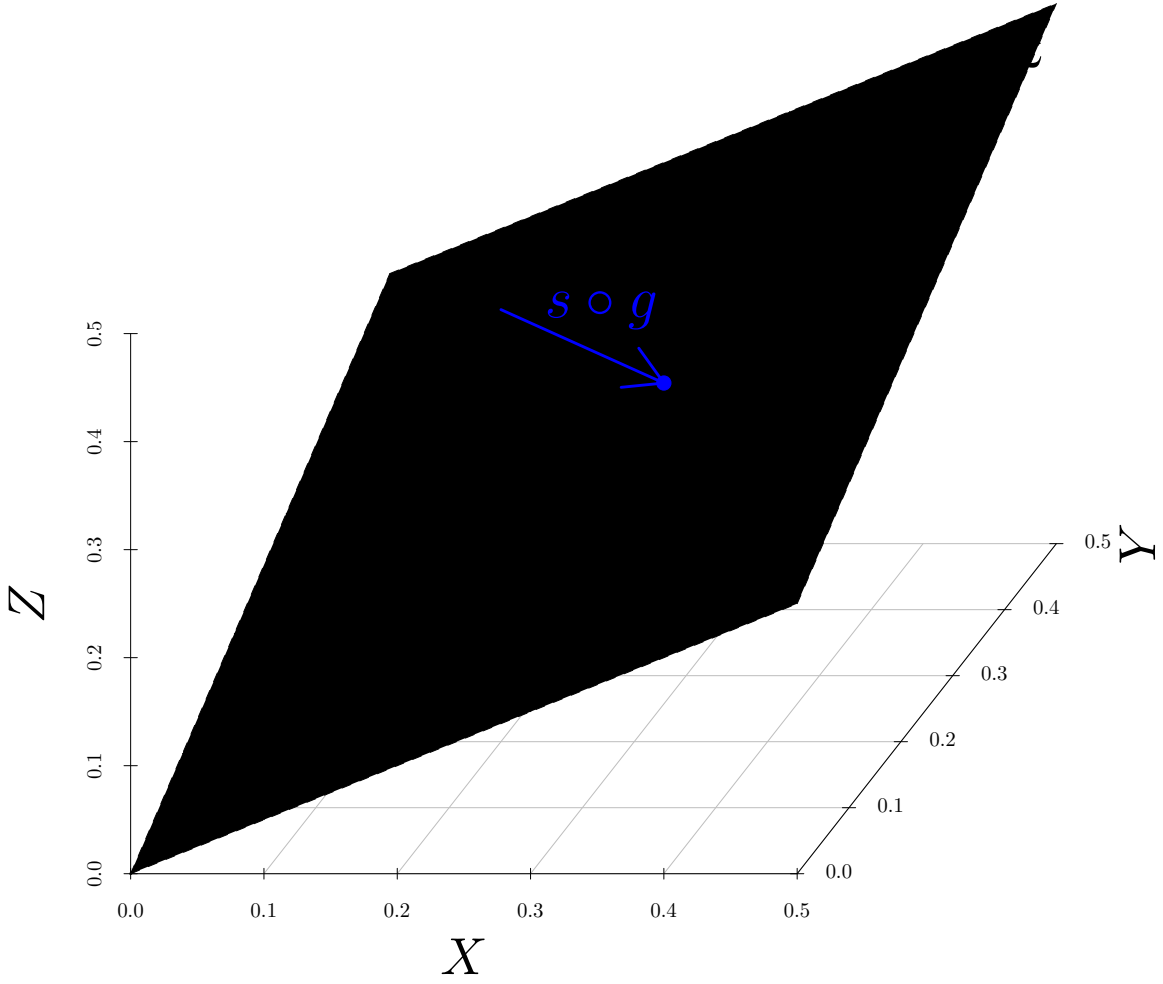


Figure 4: \mathcal{L} is a linear subspace of the coherent subspace \mathfrak{s} . If $s \circ g$ is a projection not onto \mathfrak{s} but onto \mathcal{L} , then $\boldsymbol{\mu} \in \mathfrak{s}$ will be moved to $\bar{\boldsymbol{\mu}} \in \mathcal{L}$.

3.2.1 OLS reconciliation

Recall that in OLS reconciliation, $\mathbf{R}_\perp = \mathbf{S}$ and thus it orthogonally projects $\hat{\mathbf{y}}$ to the coherent subspace. Further, it minimises the Euclidean distance between $\hat{\mathbf{y}}_{t+h|t}$ and $\tilde{\mathbf{y}}_{t+h|t}$. In addition to that Figure 5 also shows that $\tilde{\mathbf{y}}$ is always closer to \mathbf{y} than $\hat{\mathbf{y}}$ in terms of the Euclidean distance which is directly followed from the Pythagorean theorem. It also implies that the sum of squared error for OLS reconciled forecasts are always less than that for base forecasts.

3.2.2 MinT reconciliation

In MinT reconciliation, \mathbf{R}'_{\perp} is taking the form $\mathbf{S}'\mathbf{W}^{-1}$, where it can be thought of as orthogonal projections after pre-multiplying by $\mathbf{W}^{-1/2}$. That is, the coordinates of incoherent space will be scaled by $\mathbf{W}^{-1/2}$ which is then followed by the orthogonal projection. Alternatively this can be interpreted as an oblique projections in Euclidean space where the columns of \mathbf{R} is the ‘direction’ along which incoherent point forecasts are projected onto the coherent space \mathfrak{s} as depicted in Figure 3. In terms of distances, MinT minimises the Euclidean distance between $\hat{\mathbf{y}}_{t+h|t}$ and $\tilde{\mathbf{y}}_{t+h|t}$ in the transformed space which is same as the scaled Euclidean distance in the original space. Latter is also referred to as the Mahalanobis distance. We also note that the WLS is a special case of MinT where \mathbf{W}^{-1} is a diagonal matrix.

Wickramasuriya et al. (2018) showed that the MinT is optimal with respect to the mean squared forecast errors. We can provide a more general geometrical explanation to this optimality using the schematic in Figure 6. Consider the h-step ahead reconciled forecast errors. These can be always approximated by the insample h-step ahead forecast errors. Since these errors are coherent, they lies in a direction that is closer to the coherent subspace \mathfrak{s} . Therefore if you project $\hat{\mathbf{y}}$ along the direction of these in-sample forecast errors, then you can get closer to the true value \mathbf{y} as depicted in the schematic. Further, unlike OLS, the squared error for MinT reconciled forecasts is not always less than that of base forecasts in every single replication although it outperforms on average.

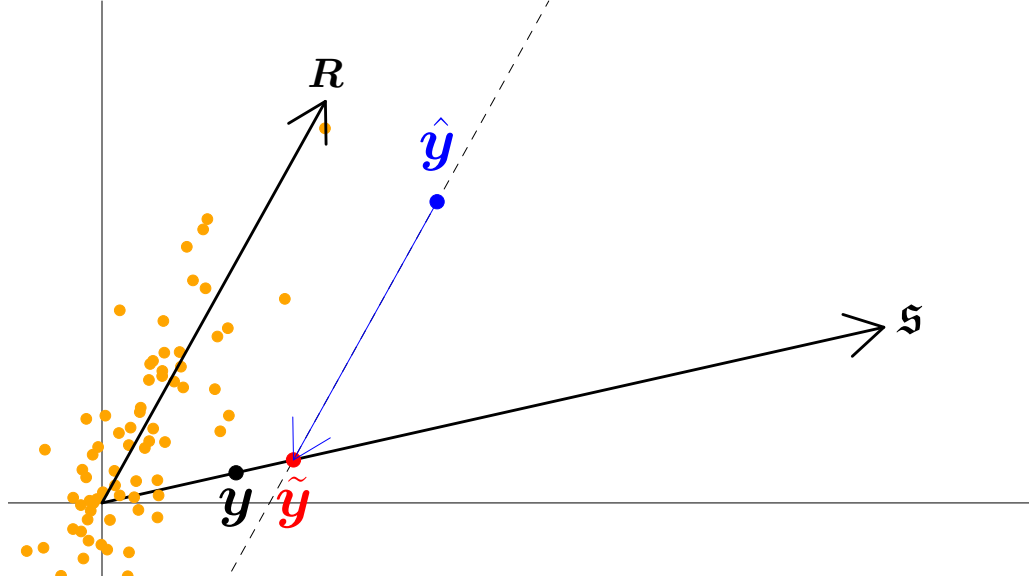


Figure 6: A schematic to represent MinT reconciliation. Points in orange colour represent the insample errors. \mathbf{R} shows the direction of the insample errors. $\hat{\mathbf{y}}$ is projected onto \mathbf{s} along the the direction of \mathbf{R} .

3.2.3 Bottom-up method

Bottom-up method is one of the traditional and simplest ways of producing coherent forecasts. Under this approach, the incoherent forecasts are projected to the coherent subspace along the direction which is perpendicular to the bottom level series. In terms of distances, this method minimises the distance between reconciled and unreconciled forecasts only

along the dimension corresponding to the bottom-level series. Therefore bottom-up methods should be thought of as a boundary case of reconciliation methods, since they ultimately do not use information at all levels of the hierarchy.

4 Bias correction

5 Application

6 Conclusions

References

- Gross, C. W. & Sohl, J. E. (1990), ‘Disaggregation methods to expedite product line forecasting’, *Journal of Forecasting* **9**(3), 233–254.
- Hunter, J. K. & Nachtergaele, B. (2001), *Applied analysis*, World Scientific Publishing Company.
- Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G. & Shang, H. L. (2011), ‘Optimal combination forecasts for hierarchical time series’, *Computational Statistics and Data Analysis* **55**(9), 2579–2589.
- Hyndman, R. J. & Athanasopoulos, G. (2018), *Forecasting: principles and practice, 2nd Edition*, OTexts.
- Wickramasuriya, S. L., Athanasopoulos, G. & Hyndman, R. J. (2018), ‘Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization’, *J American Statistical Association* . to appear.

