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

July 1, 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 $\boldsymbol{y}_t \in \mathbb{R}^n$ 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 $\boldsymbol{b}_t \in \mathbb{R}^m$ be a vector comprised of observations of all bottom-level series at time t. The hierarchical structure of the data implies that

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

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

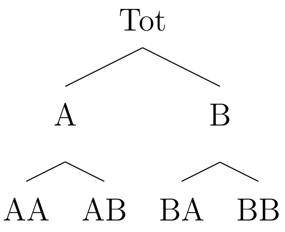


Figure 1: An example of a two level hierarchical structure.

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

$$\boldsymbol{S} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ & \multicolumn{2}{c}{\boldsymbol{I}_4} & \end{pmatrix},$$

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

While such a definition is completely serviceable, it obscures the full generality of many methodologies developed in the literature on so-called hierarchical time series. In fact, to apply concepts such as coherence and reconciliation, the data only require two important characteristics; the first is that they are multivariate, the second is that they adhere to linear constraints. Below we redefine the concepts of coherence and hierarchical time series using a geometric interpretation, before turning our attention to forecast reconciliation in Section 3.

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 holds for all $\boldsymbol{y} \in \mathfrak{s}$ 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 $\boldsymbol{S}$, i.e. $\mathfrak{s} = \text{span}(\boldsymbol{S})$. 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 '$\boldsymbol{S}$ matrix' would have columns $(1, 0, 1)'$ and $(0, 1, -1)'$. However, while there are multiple ways to define an $\boldsymbol{S}$ matrix, in all cases the columns will span the same coherent subspace, which is unique.

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 reference

include Li and Tang reference

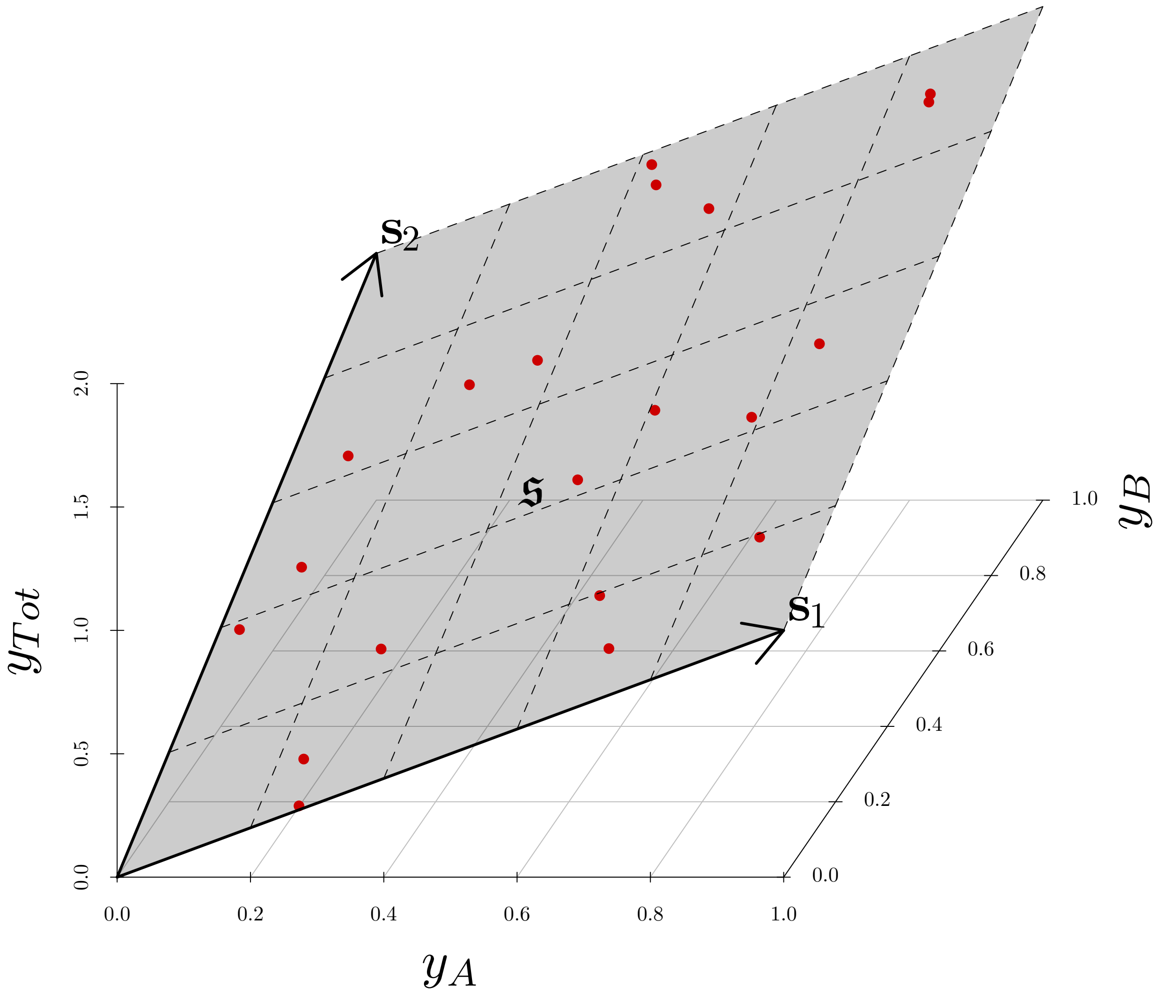


Figure 2: Depiction of a three dimensional hierarchy with $y_{\text{Tot}} = y_{\text{A}} + y_{\text{B}}$. The gray colour two dimensional plane reflects the coherent subspace $\mathfrak{s}$ where $\vec{s}_1 = (1,1,0)'$ and $\vec{s}_2 = (1,0,1)'$ are basis vectors that spans $\mathfrak{s}$. The points 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 $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T$ and all future values $\boldsymbol{y}_{T+1}, \boldsymbol{y}_{T+2}, \ldots$ lie in the coherent subspace, i.e. $\boldsymbol{y}_t \in \mathfrak{s} \quad \forall t$.

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 linear constraints are known to hold for all realisations.

include references

Definition 2.3 (Coherent Point Forecasts)**.** Let $\breve{\boldsymbol{y}}_{t+h|t} \in \mathbb{R}^n$ 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 $\breve{\boldsymbol{y}}_{t+h|t}$ is *coherent* if $\breve{\boldsymbol{y}}_{t+h|t} \in \mathfrak{s}$.

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.

some discussion about why reconciliation v single level

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 Forecast reconciliation

As discussed in the previous section, for a number of reasons, coherence is not guaranteed when forecasts are produced for all series. To ensure aligned decision making, it is desirable to adjust forecasts ex post to ensure coherence. This process is referred to as *reconciliation*. In the most general terms, reconciliation can be defined as follows

Definition 3.1 (Reconciled forecasts)**.** Let ψ be a mapping, $\psi : \mathbb{R}^n \rightarrow \mathfrak{s}$. The point forecast $\tilde{\boldsymbol{y}}_{t+h|t}$ "reconciles" $\hat{\boldsymbol{y}}_{t+h|t}$ with respect to the mapping $\psi(.)$ iff

$$\tilde{\boldsymbol{y}}_{t+h|t} = \psi\left(\hat{\boldsymbol{y}}_{t+h|t}\right) . \tag{2}$$

All reconciliation methods that we are aware of consider a linear mapping for ψ, which involves pre-multiplying base forecasts by an $n \times n$ matrix that has $\mathfrak{s}$ as its image. In , this matrix is written in the form $\boldsymbol{SG}$ (with $\boldsymbol{P}$ used in place of $\boldsymbol{G}$ in some cases), where $\boldsymbol{G}$ is an $(n - m) \times n$ matrix. This facilitates an interpretation of reconciliation as a two-step process, in the first step base forecasts $\hat{\boldsymbol{y}}_{t+h|t}$ are combined to form a new set of bottom level forecasts, in the second step, these mapped to a full vector of coherent forecasts via pre-multiplication by $\boldsymbol{S}$.

references

3.1 Unbiasedness preserving property

A sensible property is that reconciliation should not change base forecasts that are already coherent. That is

$$\tilde{\boldsymbol{y}}_{t+h|t} = \boldsymbol{SG}\hat{\boldsymbol{y}}_{t+h|t} = \hat{\boldsymbol{y}}_{t+h|t} \text{ iff } \hat{\boldsymbol{y}}_{t+h|t} \in \mathfrak{s} \, . \tag{3}$$

Using geometric intuition, it should be clear that the property in Equation 3 will hold when $\boldsymbol{SG}$ is a projection matrix . This also implies that the the property in Equation 3 does not hold for arbitrary $\boldsymbol{G}$. For this reason it is often assumed that $\boldsymbol{SGS} = \boldsymbol{S}$ or

perhaps find a reference - it is so fundamental that is on wikipedia alternatively that $\boldsymbol{GS} = \boldsymbol{I}$. Enforcing these assumptions are equivalent to assuming that $\boldsymbol{SG}$ is a projection matrix.

perhaps elaborate in a proof in appendix

The property that projections map all vectors in the coherent subspace onto themselves is useful in proving the unbiasedness preserving property of reconciliation, previously proven by. Suppose that the target of a point forecast is $\boldsymbol{\mu}_{t+h|t} := \mathrm{E}(\boldsymbol{y}_{t+h} \mid \boldsymbol{y}_1, \dots, \boldsymbol{y}_t)$ where the expectation is taken over the predictive density. Our point forecast is an estimate of this quantity that is random due to sampling variation. The point forecast will be unbiased if $\mathrm{E}_{1:t}(\hat{\boldsymbol{y}}_{t+h|t}) = \boldsymbol{\mu}_{t+h|t}$, where the subscript $1:t$ denotes an expectation taken over the training sample.

reference Shanika and maybe van Erven Culigari

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

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

$$\mathrm{E}_{1:t}(\tilde{\boldsymbol{y}}_{t+h|t}) = \mathrm{E}_{1:t}(\boldsymbol{SG}\hat{\boldsymbol{y}}_{t+h|t}) = \boldsymbol{SG}\mathrm{E}_{1:t}(\hat{\boldsymbol{y}}_{t+h|t}) = \boldsymbol{SG}\boldsymbol{\mu}_{t+h|t}.$$

Since $\boldsymbol{\mu}_{t+h|t}$ is an expectation taken with respect to the degenerate predictive density it must lie in $\mathfrak{s}$. We have already established that when $\boldsymbol{SG}$ is a projection onto $\mathfrak{s}$ then it maps all vectors in $\mathfrak{s}$ onto themselves. As such $\boldsymbol{SG}\boldsymbol{\mu}_{t+h|t} = \boldsymbol{\mu}_{t+h|t}$ when $\boldsymbol{SG}$ is a projection matrix. □

We note that the above result holds when the projection $\boldsymbol{SG}$ is only onto the coherent subspace $\mathfrak{s}$ and not for all projection matrices in general. To describe this more explicitly suppose $\boldsymbol{SG}$ has as its image $\mathfrak{L}$ which is itself a a lower dimensional linear subspace of $\mathfrak{s}$, i.e. $\mathfrak{L} \subset \mathfrak{s}$. Then for $\left\{\boldsymbol{\mu}_{t+h|t} : \boldsymbol{\mu}_{t+h|t} \in \mathfrak{s}, \boldsymbol{\mu}_{t+h|t} \notin \mathfrak{L}\right\}$, $\boldsymbol{SG}\boldsymbol{\mu}_{t+h|t} \neq \boldsymbol{\mu}_{t+h|t}$. This is depicted in Figure 3 where $\boldsymbol{\mu}_{t+h|t}$ is projected to a point $\bar{\boldsymbol{\mu}}$ in $\mathfrak{L}$. Therefore in this case, the reconciled forecast will have as its expectation $\bar{\boldsymbol{\mu}}$ rather than $\boldsymbol{\mu}_{t+h|t}$ and be biased. This result has implications in practice, in particular, the top-down method (Gross & Sohl 1990) has

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

where $\boldsymbol{p} = (p_1, \dots, p_m)'$ is an m-dimensional vector consisting a set of proportions which is use to disaggregate the top-level forecasts along the hierarchy. In this case it can be verified that $\boldsymbol{SG}$ is idempotent, i,e. $\boldsymbol{SGSG} = \boldsymbol{SG}$ and therefore $\boldsymbol{SG}$ is a projection matrix. However the image of this projection is not an m-dimensional subspace but a 1-dimensional subspace. As such, top-down reconciliation will bias base forecasts when those base forecasts are unbiased.

3.2 Why do projections work?

Now let $\boldsymbol{y}_{t+h}$ be the realisation of the data generating process at time $t+h$, and let $\|\boldsymbol{v}\|_2$ be the L_2 norm of vector $\boldsymbol{v}$. 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 $\tilde{\boldsymbol{y}}_{t+h|t} = \boldsymbol{SG}\hat{\boldsymbol{y}}_{t+h|t}$, where $\boldsymbol{G}$ is such that $\boldsymbol{SG}$ is an orthogonal projection onto $\mathfrak{s}$, then the following inequality holds:*

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

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

$$\begin{aligned} \|(\tilde{\boldsymbol{y}}_{t+h|t} - \boldsymbol{y}_{t+h})\|_2 &= \|(\boldsymbol{SG}\hat{\boldsymbol{y}}_{t+h|t} - \boldsymbol{SG}\boldsymbol{y}_{t+h})\|_2 & (6)\\ &= \|\boldsymbol{SG}(\hat{\boldsymbol{y}}_{t+h|t} - \boldsymbol{y}_{t+h})\|_2. & (7) \end{aligned}$$

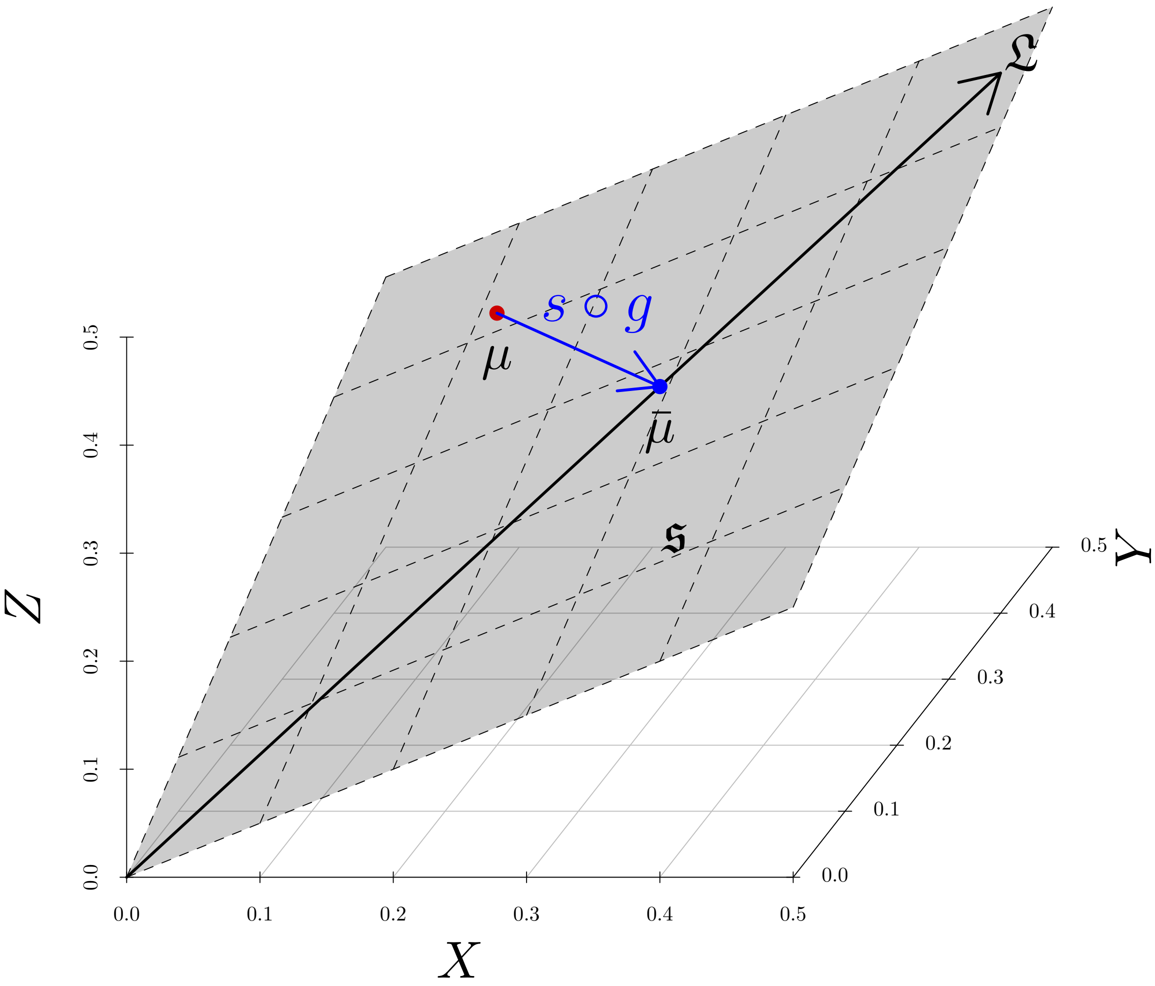


Figure 3: $\mathfrak{L}$ is a linear subspace of the coherent subspace $\mathfrak{s}$. If $s \circ g$ is a projection not onto $\mathfrak{s}$ but onto $\mathfrak{L}$, then $\boldsymbol{\mu} \in \mathfrak{s}$ will be moved to $\bar{\boldsymbol{\mu}} \in \mathfrak{L}$.

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

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

□

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

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 $\boldsymbol{R}$. Following subsections will explicitly discuss the different projection based reconciliation methods and their optimality based on distinct distance measures.

3.2.1 OLS reconciliation

Recall that in OLS reconciliation, $\boldsymbol{R}_{\perp} = \boldsymbol{S}$ and thus it orthogonally projects $\hat{\boldsymbol{y}}$ to the coherent subspace. Further, it minimises the Euclidean distance between $\hat{\boldsymbol{y}}_{t+h|t}$ and $\tilde{\boldsymbol{y}}_{t+h|t}$. In addition to that Figure 4 also shows that $\tilde{\boldsymbol{y}}$ is always closer to $\boldsymbol{y}$ than $\hat{\boldsymbol{y}}$ 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

First, the linear subspace onto which all points are projected, or the image of the projection, must be defined. In our context this can be defined by the m columns of the matrix $\boldsymbol{S}$. Second, the direction along which points are projected must be defined. This will be achieved by defining a matrix $\boldsymbol{R}$ with $n-m$ columns then span the direction of projection. A schematic of this is presented . A projection matrix can then be constructed as $\boldsymbol{S}(\boldsymbol{R}_{\perp}'\boldsymbol{S})^{-1}\boldsymbol{R}_{\perp}'$ where, $\boldsymbol{R}_{\perp}$ is an $n \times m$ orthogonal complement to $\boldsymbol{R}$ such that $\boldsymbol{R}_{\perp}'\boldsymbol{R} = \boldsymbol{0}$. It is simple to verify that this construction satisfies the properties of a projection matrix, namely symmetry and idempotence.

include

A straightforward choice of $\boldsymbol{R}$ 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 $\boldsymbol{S}$. In this case, the matrix $\boldsymbol{R}$ can be interpreted as a 'restrictions' matrix since it has the property that $\boldsymbol{R}'\boldsymbol{y} = \boldsymbol{0}$ for coherent $\boldsymbol{y}$. In OLS reconciliation, $\boldsymbol{R}_{\perp}' = \boldsymbol{S}'$ whereas in MinT or WLS reconciliation $\boldsymbol{R}_{\perp}'$ takes the form $\boldsymbol{S}'\boldsymbol{W}^{-1}$. We will be discussing these projections distinctly in the next subsection.

needs work

In MinT reconciliation, $\boldsymbol{R}_{\perp}'$ is taking the form $\boldsymbol{S}'\boldsymbol{W}^{-1}$, where it can be thought of as orthogonal projections after pre-multiplying by $\boldsymbol{W}^{-1/2}$. That is, the coordinates of incoherent space will be scaled by $\boldsymbol{W}^{-1/2}$ which is then followed by the orthogonal projection. Alternatively this can be interpreted as an oblique projections in Euclidean space where the columns of $\boldsymbol{R}$ is the 'direction' along which incoherent point forecasts are projected onto the coherent space $\mathfrak{s}$ as depicted in Figure **??**. In terms of distances, MinT minimises the Euclidean distance between $\hat{\boldsymbol{y}}_{t+h|t}$ and $\tilde{\boldsymbol{y}}_{t+h|t}$ in the transformed space which is same as the scaled Euclidean distance in the original space. Latter is also referred to as the Mahalonobis distance. We also note that the WLS is a special case of MinT where $\boldsymbol{W}^{-1}$ 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 5. 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 $\hat{\boldsymbol{y}}$ along the direction of these in-sample forecast errors, then you can get closer to the true value $\boldsymbol{y}$ 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.

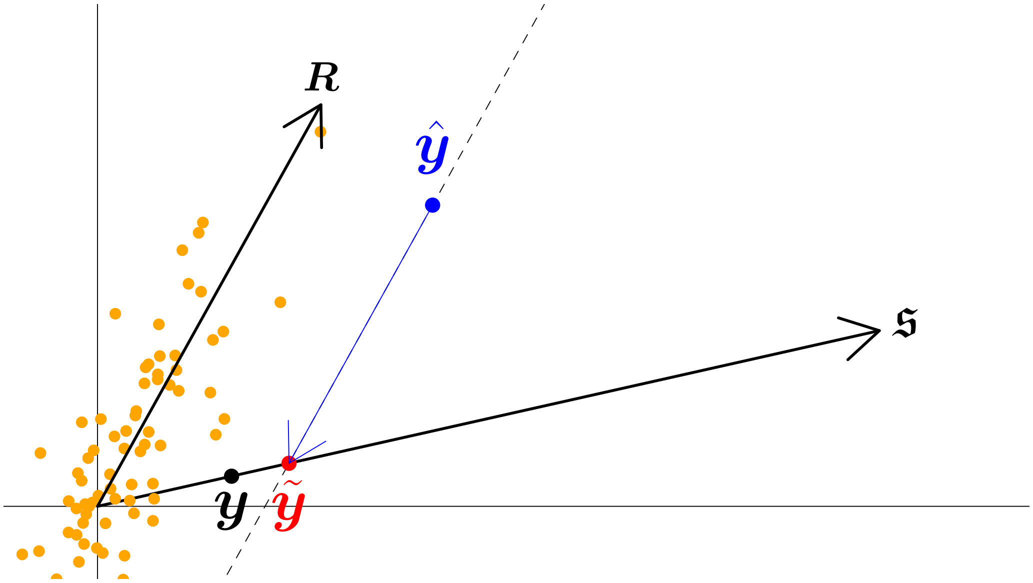


Figure 5: A schematic to represent MinT reconciliation. Points in orange colour represent the insample errors. $\boldsymbol{R}$ shows the direction of the insample errors. $\hat{\boldsymbol{y}}$ is projected onto $\mathfrak{s}$ along the the direction of $\boldsymbol{R}$.

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

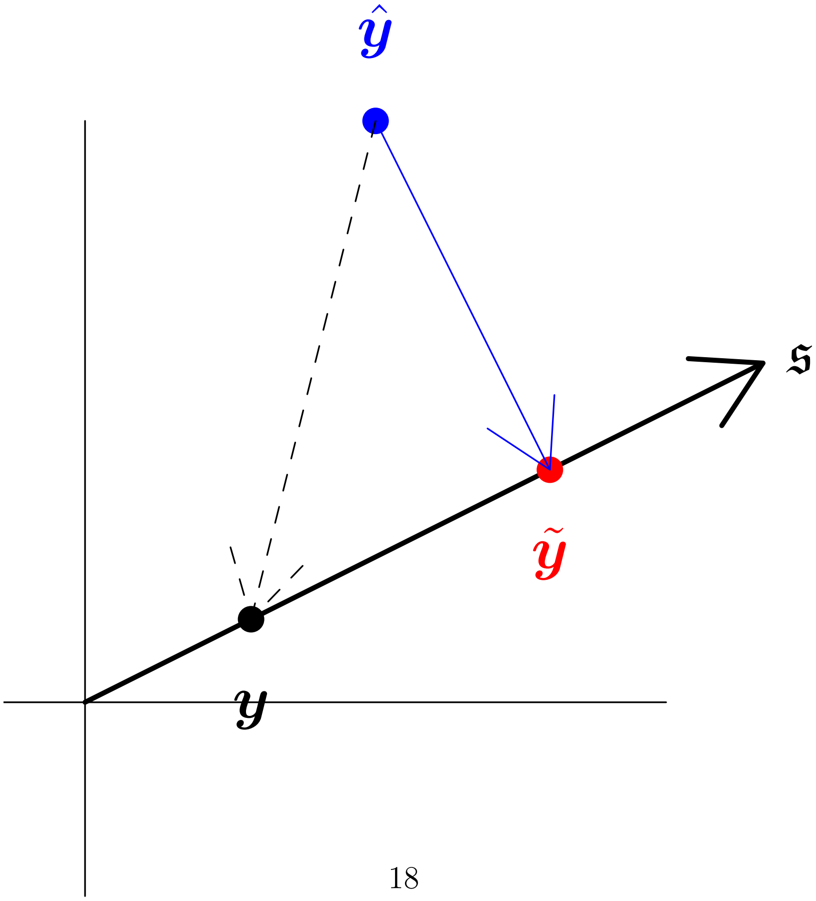
Gross, C. W. & Sohl, J. E. (1990), ‘Disaggregation methods to expedite product line forecasting’, *Journal of Forecasting* **9**(3), 233–254.

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.


$\hat{\boldsymbol{y}}$
$\mathfrak{s}$
$\tilde{\boldsymbol{y}}$
$\boldsymbol{y}$