Hierarchical Forecasting

Name of First Author and Name of Second Author

1 Introduction

TBC

Name of First Author
Name, Address of Institute, e-mail: name@email.address

Name of Second Author
Name, Address of Institute e-mail: name@email.address

2 Coherent Point forecasts

Due to its importance in many research applications, coherency is having a well established literature in terms of hierarchical point forecasts. In this section we will review these traditional coherent forecasting methods. First let us start with the notations.

2.1 Notations and preliminaries

We follow the notations introduced by Wickramasuriya et al. (2018) and Gamakumara et al. (2018) where necessary. Suppose $\mathbf{y}_t \in \mathbb{R}^n$ comprises all observations of the hierarchy at time t and $\mathbf{b}_t \in \mathbb{R}^m$ comprises only the bottom level observations at time t. Then due to the aggregation nature of the hierarchy we have

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

where $\mathbf{S}$ is an $n \times m$ constant matrix whose columns span the linear subspace for which all constraints hold. $\mathbf{S}$ is also referred to as the "summing matrix" since it aggregates the observations of bottom level to the corresponding upper levels. We can also think, pre-multiplying by $\mathbf{S}$ will map a vector in $\mathbb{R}^m$ to a vector in $\mathbb{R}^n$.

In any hierarchy, the most aggregated level is labelled level 0, the second most aggregated level is labelled level 1 and so on to the most disaggregate level k.

Consider the hierarchy given in Figure **??**. This example consists of two levels. At a particular time t, let $y_{Tot,t}$ denote the observation at level 0; $y_{A,t}, y_{B,t}$ denote observations at level 1; and $y_{AA,t}, y_{AB,t}, y_{BA,t}, y_{BB,t}$ denote observations at level 2. Then $\mathbf{y}_t = [y_{Tot,t}, y_{A,t}, y_{B,t}, y_{C,t}, y_{AA,t}, y_{AB,t}, y_{BA,t}, y_{BB,t}]'$, $\mathbf{b}_t = [y_{AA,t}, y_{AB,t}, y_{BA,t}, y_{BB,t}]'$, $m = 4$, $n = 7$, and

$$\mathbf{S} = \begin{pmatrix} 1\ 1\ 1\ 1 \\ 1\ 1\ 0\ 0 \\ 0\ 0\ 1\ 1 \\ I_4 \end{pmatrix}, \tag{2}$$

where $\mathbf{I}_4$ is a 4-dimension identity matrix.

2.2 Coherent forecasts

Various approaches have been used in literature to produce coherent forecasts of hierarchical time series incorporating valuable, structural information of the hierarchy. Let us first start with the novel definition given by Gamakumara et al. (2018). While the coherency is conceptually illustrated in literature many times, this definition provides a geometrical understanding to the problem which also facilitate extension to the situation of probabilistic forecasting discuss in section 4.

Definition 1 (Coherent subspace).
The m-dimensional linear subspace $\mathfrak{s} \subset \mathbb{R}^n$ that is spanned by the columns of $\mathbf{S}$, i.e. $\mathfrak{s} = \text{span}(\mathbf{S})$, is defined as the *coherent space*.

For a particular coherent subspace $\mathfrak{s}$, there exist several distinct basis vectors. For example, in the smallest hierarchy with two bottom level series, (A,B) that add up to the top level (Tot), $\{(1\ 1\ 0)', (1\ 0\ 1)'\}$, $\{(1\ 0\ 1)', (0\ 1\ -1)'\}$ are alternative basis vectors that span the same $\mathfrak{s}$. Moreover, the singular value decomposition of these vectors are some other basis vectors for the same. Given a basis for $\mathfrak{s}$, every series of the hierarchy can be linearly determined as a linear combination of those basis vectors. We refer to the coefficients of these linear combinations as the *basis series*. It is apparent that these basis series are m-dimensional and linearly independent in a given hierarchy. For example, in the smallest hierarchy, $(y_{A,t}, y_{B,t})$ and $(y_{Tot,t}, y_{A,t})$ are the basis series corresponding to the basis vectors $\{(1\ 1\ 0)', (1\ 0\ 1)'\}$ and $\{(1\ 0\ 1)', (0\ 1\ -1)'\}$ respectively. Further, bottom level series is a basis series that corresponds to the column vectors of $\mathbf{S}$.

Because the basis is not unique for a given coherent subspace, the following definitions are not unique, and one can redefine them with respect to any basis. However, we consider the basis defined by the columns of $\mathbf{S}$ in what follows.

Definition 2 (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 $\check{\mathbf{y}}_{t+h|t} \in \mathfrak{s}$.

To understand this definition more clearly, let us consider the smallest hierarchy. Suppose the forecasts of these series at time $t+h$ are given by $\check{\mathbf{y}}_{t+h} = [\check{y}_{Tot,t+h}, \check{y}_{A,t+h}, \check{y}_{B,t+h}]$. Due to the aggregation constraint of the hierarchy we have $\check{y}_{Tot,t+h} = \check{y}_{A,t+h} + \check{y}_{B,t+h}$. This implies that, even though $\check{\mathbf{y}}_{t+h} \in \mathbb{R}^3$, the points actually lie in $\mathfrak{s} \subset \mathbb{R}^3$, which is a two dimensional subspace within $\mathbb{R}^3$ space.

Although the formal definition for coherent point forecasts is formed recently, it is intuitively used in many studies in literature. First let us explore the most traditional approaches. Generally, these involve forecasting one level of aggregation and aggregate them up or disaggregate them according to the level they were chosen.

2.2.1 Bottom-up approach

In bottom-up approach, forecasts of the lowest level series are first generated and these are aggregated to get the forecasts at upper levels of the hierarchy (Dunn et al. (1976)). That is, considering the hierarchy given in **??**, first the forecasts for AA, AB, BA and BB are obtained. Then the forecasts of series A is obtained by simply adding the forecasts of AA and AB. Similarly the forecasts of B is obtained by adding the forecasts of BA and BB. The forecasts of Tot series is then the aggregation of A and B which in turn the sum of forecasts of AA, AB, BA and BB.

In terms of notations, let $\hat{\mathbf{b}}_{t+h|t} \in \mathbb{R}^m$ consists h-step ahead forecasts of the bottom levels series. i.e. $\hat{\mathbf{b}}_{t+h|t} = (\hat{y}_{AA,t+h|t}, \hat{y}_{AB,t+h|t}, \hat{y}_{BA,t+h|t}, \hat{y}_{BB,t+h|t})$, where, $\hat{y}_{i,t+h|t}$ is the h-step ahead forecasts of i^{th} series. Then, the bottom-up forecasts are given by,

$$\breve{\mathbf{y}}^{BU}_{t+h|t} = \mathbf{S}\hat{\mathbf{b}}_{\mathbf{t+h|t}}. \tag{3}$$

Even though this method looks easy to construct, it is less reliable. In fact, bottom up approach provides accurate forecasts only if the bottom level series of the hierarchy are accurately forecast. However, if the bottom level series are highly volatile or too noisy, they are challenging to forecast. Then the bottom-up approach would produce inaccurate point forecasts.

2.2.2 Top-down approach

In contrast to the bottom-up method, the top-down approach involves forecasting the most aggregated series first and then disaggregating these forecasts down the hierarchy.

In general, all top-down forecasts can be written as,

$$\breve{\mathbf{y}}^{TD}_{t+h|t} = \mathbf{S}\hat{y}_{t+h|t}\mathbf{p}, \tag{4}$$

for $j = 1, ..., m$ where $\mathbf{p}$ is a vector consisting the disaggregation proportions of the bottom level series with respect to the top level series.

Usually, these proportions are calculated based on observed data, which is referred to as historical proportions. Gross & Sohl (1990) provide a comprehensive summary on using historical proportions in this context. Commonly there are two approaches of using historical proportions. One is to use the average of historical proportions of desired bottom level series relative to the total series. That is, the historical proportion of j^{th} series is given by,

$$p_j = \frac{1}{T}\sum_{t=1}^{T}\frac{Y_{j,t}}{Y_{tot,t}}. \tag{5}$$

The second approach is to use the proportion of average of the bottom level series $y_{j,t}$ over the time $t = 1, ..., T$ relative to that of the total series $y_{Tot,t}$. i.e.

$$p_j = \frac{\frac{1}{T}\sum_{t=1}^{T} y_{j,t}}{\frac{1}{T}\sum_{t=1}^{T} y_{Tot,t}} \tag{6}$$

However, the largest limitation of these approaches is that its inability to reflect the characteristics of individual series such as trends, seasonality or other special events, in the forecasts of disaggregate levels. This will mainly effect the hierarchies with series having different patterns in different levels.

To overcome this limitation, Athanasopoulos et al. (2009) introduced a new top-down approach which disaggregate the top level forecasts according to the propor-

tions of forecasts rather than historical proportions. They found that their method outperforms the conventional top-down approach through an empirical application. To discuss this approach in detail, we refer to their notations as follows. Suppose, $y_{j,T+h|T}^{(i)}$ denote the h-step ahead forecasts series in i levels above j. Further suppose $\Sigma(\hat{y}_{i,T+h|T})$ denote the h-step ahead sum of forecast of all child nodes corresponds to the parent node i. Then,

$$p_j = \prod_{i=0}^{K-1} \frac{\hat{y}_{j,T+h|T}^{(i)}}{\Sigma(\hat{y}_{j,T+h|T}^{(i+1)})}, \tag{7}$$

for $j = 1, ..., m$.

For the hierarchy given in **??**, the proportions of forecasts are calculated as follows.

$$p_1 = \left(\frac{\hat{y}_{AA,T+h|T}}{\hat{y}_{AA,T+h|T} + \hat{y}_{AB,T+h|T}}\right)\left(\frac{\hat{y}_{A,T+h|T}}{\hat{y}_{A,T+h|T} + \hat{y}_{B,T+h|T}}\right), \tag{8}$$

$$p_2 = \left(\frac{\hat{y}_{AB,T+h|T}}{\hat{y}_{AA,T+h|T} + \hat{y}_{AB,T+h|T}}\right)\left(\frac{\hat{y}_{A,T+h|T}}{\hat{y}_{A,T+h|T} + \hat{y}_{B,T+h|T}}\right), \tag{9}$$

$$p_3 = \left(\frac{\hat{y}_{BA,T+h|T}}{\hat{y}_{BA,T+h|T} + \hat{y}_{BB,T+h|T}}\right)\left(\frac{\hat{y}_{B,T+h|T}}{\hat{y}_{A,T+h|T} + \hat{y}_{B,T+h|T}}\right), \tag{10}$$

$$p_4 = \left(\frac{\hat{y}_{BB,T+h|T}}{\hat{y}_{BA,T+h|T} + \hat{y}_{BB,T+h|T}}\right)\left(\frac{\hat{y}_{B,T+h|T}}{\hat{y}_{A,T+h|T} + \hat{y}_{B,T+h|T}}\right). \tag{11}$$

Recently, Mirčetić et al. (2017) proposed a modified approach to the conventional top-down methods, which involves forecasting h-step ahead proportions of bottom level series relative to the top level series. That is, let the ratio between the j^{th} bottom level series and the top level series over the period of $t = 1, ..., T$ is given by, $p_{j,t} = y_{j,t}/y_{Tot,t}$. Then this series of proportions will be forecast for the h-step ahead period, which is denoted by $\hat{p}_{j,T+h|T}$. These proportions will be then used as the disaggregate proportions to construct coherent top-down forecasts. Thus giving,

$$p_j = \hat{p}_{j,T+h|T} \tag{12}$$

This approach is much reliable, since it calculates the proportions as a function time, rather than using the simple averages in traditional top-down methods.

However, the common limitation of all these top-down approaches is that they will not produce unbiased forecasts even if the base forecasts are unbiased. We will discuss about the unbiased property in the following sections.

2.2.3 Middle-out approach

A compromise between these two approaches is the middle-out method which entails forecasting each series of a selected middle level in the hierarchy and then

forecasting upper levels by the bottom-up method and lower levels by the top-down method.
Since this is a hybrid approach of bottom-up and top-down approaches middle-out method still remains the limitations of those two up to a certain extent.

3 Point forecast reconciliation

The common limitation of traditional hierarchical forecasting methods is that the loss of structural information of individual series, as it precisely forecasting only one level of series in the hierarchy.

As an alternative to these traditional methods, Hyndman et al. (2011) proposed to utilize the information from all levels of the hierarchy to obtain coherent point forecasts in a two stage process. In the first stage, the forecasts of all series are independently obtained by fitting univariate models for individual series in the hierarchy. It is very unlikely that these forecasts are coherent. Thus in the second stage, these forecasts are optimally combined through a regression model to obtain coherent forecasts. This second step is referred to as "reconciliation" since it takes a set of incoherent forecasts and revises them to be coherent.

This is the very first study that introduced the concept of forecast reconciliation to the hierarchical literature. Since this approach starts from the individual forecasts, it uses all the relevant information from the hierarchy in producing coherent forecasts. Thus improves the drawbacks from traditional approaches, in particularly the loss of information.

Hierarchical point forecast reconciliation is broadly studied and implemented in many research applications. However a proper definition for this important concept is recently given by Gamakumara et al. (2018) as stated below.

Suppose we fit univariate time series models for all series based on data upto time t and obtained h-step ahead forecasts. These forecasts are then formed in a vector $\hat{\mathbf{y}}_{T+h}$ by stacking the forecasts at each node in the same order as $\mathbf{y}_t$. Since these do not satisfy the aggregate constraints of the hierarchy, they are often referred to as "incoherent" forecasts. In some texts these also referred to as "base" forecasts. Further let $\mathbf{G}$ and $\mathbf{d}$ be an $m \times n$ matrix and $m \times 1$ vector respectively, and let $g : \mathbb{R}^n \to \mathbb{R}^m$ be the mapping $g(\mathbf{y}) = \mathbf{G}\mathbf{y} + \mathbf{d}$. A composition of g and $s(.)$ gives the following definition for point forecast reconciliation.

Definition 3.
The point forecast $\tilde{\mathbf{y}}_{t+h|t}$ "reconciles" $\hat{\mathbf{y}}_{t+h|t}$ with respect to the mapping $g(.)$ iff

$$\tilde{\mathbf{y}}_{t+h|t} = \mathbf{S}(\mathbf{G}\hat{\mathbf{y}}_{t+h|t} + \mathbf{d}). \tag{13}$$

It is worth noticing that the mapping g is a linear function and it converts unreconciled forecasts into new bottom level forecasts. Subsequently, pre-multiplication

by $\mathbf{S}$ will linearly project these onto the coherent subspace $\mathfrak{s}$ and thus giving coherent point forecasts following definition (2). One could also consider a non-linear function for $g(.)$ which will then perform a non-linear reconciliation. This is a possible extension of hierarchical forecasting that we leave space for later discussion under a different topic.

Previous studies in hierarchical point forecasting have only focussed on the linear case, $\mathbf{g}(.) = \mathbf{G}\hat{\mathbf{y}}$, where $\mathbf{G}$ is an $m \times n$ matrix and $\mathbf{d} = \mathbf{0}$, so $\tilde{\mathbf{y}}_{t+h} = \mathbf{S}\mathbf{G}\hat{\mathbf{y}}_{t+h}$. It should also worth noticing that in almost all past studies have used the notation $\mathbf{P}$ for which we use $\mathbf{G}$.

Let us now illustrate the definition 3 and observe the structure of matrix $\mathbf{G}$. Let $\mathbf{R} \in \mathbb{R}^{n\times(n-m)}$ comprise the columns that span the null space of $\mathfrak{s}$. Note that $\mathbf{R}$ is not unique; one example is a matrix whose columns represent the aggregation constraints for a given hierarchy. For the simplest hierarchy with two series add up to the top,

$$\mathbf{S} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{R} = \begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix}.$$

Further let $\{\mathbf{s}_1, \ldots, \mathbf{s}_m\}$ and $\{\mathbf{r}_1, \ldots, \mathbf{r}_{n-m}\}$ denote the columns of $\mathbf{S}$ and $\mathbf{R}$ respectively. Then $\mathbf{B} = \{\mathbf{s}_1, \ldots, \mathbf{s}_m, \mathbf{r}_1, \ldots, \mathbf{r}_{n-m}\}$ is a basis for $\mathbb{R}^n$. Now, using the insights of Definition 3, we can use the following steps to reconcile the point forecasts.

Step 1: Obtaining reconciled bottom level point forecasts

For a given incoherent set of point forecasts $\hat{\mathbf{y}}_{t+h} \in \mathbb{R}^n$, first we find the coordinates of $\hat{\mathbf{y}}_{t+h}$ with respect to the basis $\mathbf{B}$. Let $(\tilde{\mathbf{b}}'_{t+h}, \tilde{\mathbf{a}}'_{t+h})'$ denote these coordinates. Note that $\tilde{\mathbf{b}}_{t+h}$ is a basis series which is equivalent to the reconciled bottom level series, and corresponds to the coordinates of the basis $\{\mathbf{s}_1, \ldots, \mathbf{s}_m\}$. Similarly, $\tilde{\mathbf{a}}_{t+h}$ is another basis series corresponding to the coordinates of the basis $\{\mathbf{r}_1, \ldots, \mathbf{r}_{n-m}\}$. Then from basic properties of linear algebra it follows that,

$$(\mathbf{S} \;\; \mathbf{R})(\tilde{\mathbf{b}}'_{t+h}, \; \tilde{\mathbf{a}}'_{t+h})' = \hat{\mathbf{y}}_{t+h},$$

$$\hat{\mathbf{y}}_{t+h} = \mathbf{S}\tilde{\mathbf{b}}_{t+h} + \mathbf{R}\tilde{\mathbf{a}}_{t+h}, \tag{14}$$

and

$$(\tilde{\mathbf{b}}'_{t+h}, \; \tilde{\mathbf{a}}'_{t+h})' = (\mathbf{S} \;\; \mathbf{R})^{-1}\hat{\mathbf{y}}_{t+h}. \tag{15}$$

In order to find $(\mathbf{S} \;\; \mathbf{R})^{-1}$, let $\mathbf{S}_\perp$ and $\mathbf{R}_\perp$ be the orthogonal complements of $\mathbf{S}$ and $\mathbf{R}$ respectively. Then $(\mathbf{S} \;\; \mathbf{R})^{-1}$ is given by,

$$(\mathbf{S}\ \ \mathbf{R})^{-1} = \begin{pmatrix} (\mathbf{R}'_\perp \mathbf{S})^{-1}\mathbf{R}'_\perp \\ (\mathbf{S}'_\perp \mathbf{R})^{-1}\mathbf{S}'_\perp \end{pmatrix}. \tag{16}$$

Thus we have,

$$\begin{pmatrix} \tilde{\mathbf{b}}_{t+h} \\ \tilde{\mathbf{a}}_{t+h} \end{pmatrix} = \begin{pmatrix} (\mathbf{R}'_\perp \mathbf{S})^{-1}\mathbf{R}'_\perp \\ (\mathbf{S}'_\perp \mathbf{R})^{-1}\mathbf{S}'_\perp \end{pmatrix} \hat{\mathbf{y}}_{t+h}. \tag{17}$$

From (**??**) it follows that,

$$\tilde{\mathbf{b}}_{t+h} = (\mathbf{R}'_\perp \mathbf{S})^{-1}\mathbf{R}'_\perp \hat{\mathbf{y}}_{t+h} \tag{18}$$

Step 2: Obtaining reconciled point forecasts for the whole hierarchy

This step directly follows from the definition for coherent forecasts. To obtain reconciled point forecasts for the entire hierarchy, we map $\tilde{\mathbf{b}}_{t+h} \in \mathbb{R}^n$ to the $\mathfrak{s}$ through $\mathbf{S}$. Thus we have,

$$\tilde{\mathbf{y}}_{t+h} = \mathbf{S}(\mathbf{R}'_\perp \mathbf{S})^{-1}\mathbf{R}'_\perp \hat{\mathbf{y}}_{t+h}, \qquad \tilde{\mathbf{y}}_{t+h} \in \mathfrak{s}, \tag{19}$$

and the $\mathbf{G}$ we defined before is having the structure

$$\mathbf{G} = (\mathbf{R}'_\perp \mathbf{S})^{-1}\mathbf{R}'_\perp \tag{20}$$

Finding a suitable $\mathbf{R}_\perp$ with respect to a certain loss function will lead to optimally reconciled point forecasts of the hierarchy.

It is also worth discussing the unbiased property of these reconciled forecasts. Let us assume the incoherent forecasts, $\hat{\mathbf{y}}_{t+h}$ are unbiased. That is, $E_{1:t}(\hat{\mathbf{y}}_{t+h|t}) = \mu_{t+h|t}$ where $\mu_{t+h|t} = E_{1:t}[\mathbf{y}_{t+h}|\mathbf{y}_1, ..., \mathbf{y}_t]$ is the true mean of the hierarchy. Here the expectation is taken over the observed training set. Then for any $\mathbf{G}$ such that $\mathbf{SGS} = \mathbf{S}$ produce unbiased reconciled forecasts (Hyndman et al. (2011)). Further Gamakumara et al. (2018) showed that reconciled point forecasts are unbiased only if $s \circ g$ is a projection through the theorem they called *unbiasedness preserving property*. This implies that projection is playing an important role in producing unbiased reconciled forecasts.

The bottom-up method is also producing unbiased forecasts since it is projecting the incoherent forecasts to the coherent subspace along the perpendicular direction of the dimension corresponds to the bottom-levels. Thus the bottom-up approach can be considered as the boundary case of reconciliation methods. However, it is not always a preferred approach since it uses only a part of the information available in the hierarchy. Further as shown by Hyndman et al. (2011), any top-down approach is not producing unbiased coherent forecasts even if the top level base forecasts are unbiased, as it adding a bias component to each disaggregate level.

Now let us discuss on how the previous reconciliation methods were constructed in detail and how they correspond with the above reconciliation arguments.

3.1 OLS

Intuitively, the reconciliation starts with a set of incoherent forecasts, $\hat{\mathbf{y}}_{T+h}$. Hyndman et al. (2011) proposed to optimally combine these incoherent forecasts through the following regression model.

$$\hat{\mathbf{y}}_{t+h} = \mathbf{S}\beta_{t+h} + \varepsilon_{t+h}, \tag{21}$$

where $\beta_{t+h} = E[\mathbf{b}_{t+h}|\mathbf{b}_1, ..., \mathbf{b}_t]$ is the unknown mean of the bottom level series at time $t+h$ and ε_{t+h} is the reconciliation error with mean zero and variance $\mathbf{V}$. The ordinary least squares (OLS) solution for the above regression model gives,

$$\hat{\beta}_{t+h} = (\mathbf{S}'\mathbf{S})^{-1}\mathbf{S}'\hat{\mathbf{y}}_{t+h}, \tag{22}$$

and thus giving the reconciled forecasts,

$$\tilde{\mathbf{y}}_{t+h} = \mathbf{S}\hat{\beta}_{t+h} = \mathbf{S}(\mathbf{S}'\mathbf{S})^{-1}\mathbf{S}'\hat{\mathbf{y}}_{t+h}. \tag{23}$$

Recall that the $\mathbf{G}$ matrix is having a structure $\mathbf{G} = (\mathbf{R}'_{\perp}\mathbf{S})^{-1}\mathbf{R}'_{\perp}$ according to equation (19). Thus when $\mathbf{R}'_{\perp} = \mathbf{S}'$, it corresponds to the OLS reconciliation. In OLS reconciliation, the incoherent forecasts are orthogonally projected to the coherent subspace $\mathfrak{s}$. Thus it will minimises the Euclidean distance between $\hat{\mathbf{y}}$ and $\tilde{\mathbf{y}}$.

In a later study by Wickramasuriya et al. (2018) showed that the variance covariance matrix of the reconciliation error in the above regression model is not identifiable and thus the GLS solution for the same model is unattainable.

3.2 MinT

As an improved work of Hyndman et al. (2011), Wickramasuriya et al. (2018) find a unique analytical solution for the reconciled point forecasts which are unbiased, by minimizing the sum of variances of reconciled forecast errors. An important feature of this method is that it imposes the correlation structure of the whole hierarchy to produce reconciled point forecasts. Their simulation study illustrates that this approach outperforms all existing hierarchical forecasting methods.

Suppose the variance of h-step ahead base forecast errors is denoted by, $Var(\mathbf{y}_{T+h} - \hat{\mathbf{y}}_{T+h}) = \mathbf{W}_{T+h}$. Wickramasuriya et al. (2018) first showed that the variance of the reconciled forecast errors, i.e $Var(\mathbf{y}_{T+h} - \tilde{\mathbf{y}}_{T+h}) = \tilde{\mathbf{W}}_{T+h}$ is given by,

$$\tilde{\mathbf{W}}_{T+h} = \mathbf{S}\mathbf{G}\mathbf{W}_{T+h}\mathbf{G}'\mathbf{S}' \tag{24}$$

for any choice of $\mathbf{G}$. Then they minimize the trace of $\tilde{\mathbf{W}}_{T+h}$ with respect to $\mathbf{SGS} = \mathbf{S}$ to obtain optimal $\mathbf{G}$. By imposing the unbiased constraint they ensure that the resulting reconciled forecasts are unbiased. The closed form solution to this minimization problem is given by,

$$\mathbf{G} = (\mathbf{S}'\mathbf{W}_{T+h}^{-1}\mathbf{S})^{-1}\mathbf{S}'\mathbf{W}_{T+h}^{-1}, \tag{25}$$

and they named this as the MinT approach. Therefore, the reconciled point forecasts are,

$$\tilde{\mathbf{y}}_{T+h} = \mathbf{SG}\hat{\mathbf{y}}_{T+h} = \mathbf{S}(\mathbf{S}'\mathbf{W}_{T+h}^{-1}\mathbf{S})^{-1}\mathbf{S}'\mathbf{W}_{T+h}^{-1}\hat{\mathbf{y}}_{T+h}. \tag{26}$$

Recall the structure of matrix $\mathbf{G}$ in equation (19) derived from the definition (3). When $\mathbf{R}_{\perp}' = \mathbf{S}'\mathbf{W}^{-1}$ this coincides with the MinT reconciliation. In terms of projections, MinT is doing an oblique projection of $\hat{\mathbf{y}}$ onto the coherent subspace $\mathfrak{s}$ along the direction of $\mathbf{R}$. Further in terms of distances, MinT minimises the Mahalonobis distance between $\hat{\mathbf{y}}$ and $\tilde{\mathbf{y}}$ as proven by Wickramasuriya et al. (2018).

Wickramasuriya et al. (2018) further discussed alternative ways to estimate $\mathbf{W}_{T+h}$ and how these estimates lead to different $\mathbf{G}$ matrices. Since the variance covariance matrix of 1-step ahead incoherent forecast errors are approximately proportional to that of h-step ahead incoherent forecasts errors, we have $\mathbf{W}_{T+h} = \alpha_{T+h}\mathbf{W}_{T+1}$ where $\alpha_{T+h} > 0$. Therefore, to estimate $\mathbf{W}_{T+h}$ it is first required to estimate $\mathbf{W}_{T+1}$.

Simply, if $\mathbf{W}_{T+1}$ is approximated by an identity matrix, then $\mathbf{G}$ in (25) collapses to OLS reconciliation. Following are a few other estimates for $\mathbf{W}_{T+1}$.

3.2.1 MinT(Sample)

The unbiased sample variance covariance matrix of 1-step ahead incoherent forecast errors can be used as the most reasonable estimator for $\mathbf{W}_{T+1}$. That is,

$$\hat{\mathbf{W}}_{T+1} = \frac{1}{T}\sum_{t=1}^{T}\hat{\mathbf{e}}_{t+1}\hat{\mathbf{e}}_{t+1}',$$

where $\hat{\mathbf{e}}_{t+1}$ consists the 1-step ahead insample forecast errors.

The resulting $\mathbf{G}$ matrix is referred to as MinT(Sample). This estimator is performing well in small hierarchies. However for the hierarchies with large dimensions, especially when $m > T$, this causes singularity problems.

3.2.2 MinT(Shrink)

Wickramasuriya et al. (2018) also proposed an shrinkage estimator for $\mathbf{W}_{T+1}$, which is apparently useful in large dimensional hierarchies. This estimator is given by,

$$\hat{\mathbf{W}}^{shr}_{T+1} = \tau\hat{\mathbf{W}}^{D}_{T+1} + (1-\tau)\hat{\mathbf{W}}_{T+1}, \tag{27}$$

where $\hat{\mathbf{W}}^{D}_{T+1}$ is a diagonal matrix comprising diagonal entries of $\hat{\mathbf{W}}_{T+1}$ and

$$\tau = \frac{\sum_{i\neq j}\hat{Var}(\hat{r}_{ij})}{\sum_{i\neq j}\hat{r}^2_{ij}}$$

is a shrinkage parameter proposed by Schäfer & Strimmer (2005) where, $\hat{r}_{ij}$ is the ij-th element of sample correlation matrix. In this estimation, the off-diagonal elements of 1-step ahead sample covariance matrix will be shrunk to zero depending on the sparsity.

3.3 WLS

If $\mathbf{W}_{T+1}$ is approximated by a diagonal matrix with diagonal elements being the variances of incoherent forecast errors, then the resulting reconciliation is referred to as WLS reconciliation. This was first proposed by Hyndman et al. (2016) with reference to the regression model introduced in Hyndman et al. (2011). However a proper theoretical justification for the WLS reconciliation is given in MinT solution.

3.3.1 GTOP

van Erven & Cugliari (2014) proposed to use a game theoretically approach to produce reconciled forecasts that are at least as good as incoherent forecasts. Initially starting with the best possible incoherent forecasts, they show that the sum of weighted squared loss of these incoherent forecasts will never be less than that for the reconciled forecasts.

This approach does not require assumptions like unbiasedness on incoherent forecasts and it allows forecasters to choose the best possible incoherent forecasts without depending on any assumptions. However it also doesn't guarantee to produce unbiased reconciled forecasts. Further, there is no closed form solution to this approach and thus would result a high computational cost in producing reconciled forecasts for large scale hierarchies.

Any reconciliation approach via projections including the MinT approach us preferred as they improved these limitations in the GTOP method. Nevertheless, they guarantee to produce coherent forecasts that are at least as good as incoherent forecasts (Wickramasuriya et al. (2018), Gamakumara et al. (2018)).

Another important feature of hierarchical time series is that they often consist of thousands or millions of individual series and this imposes computational challenges in implementing any forecasting solutions. The MinT reconciliation method was further generalized to handle these constraints and to scale to large hierarchies

by Wickramasuriya et al. (2018).

4 Probabilistic hierarchical forecasts

Point forecasts are limited because they provide no indication of forecast uncertainty. Providing prediction intervals helps, but a richer description of forecast uncertainty is obtained by estimating the entire forecast distribution. These are often called "probabilistic forecasts" (Gneiting & Katzfuss 2014). For example, McSharry et al. (2005) produced probabilistic forecasts for electricity demand, Ben Taieb et al. (2017) for smart meter data, Pinson et al. (2009) for wind power generation, and Gel et al. (2004), Gneiting et al. (2005) and Gneiting & Raftery (2005) for various weather variables.

Although there is a rich and growing literature in producing coherent point forecasts of hierarchical time series, a little attention has been given to coherent probabilistic forecasts. One relevant paper we are aware of is Taieb et al. (2017), who recently proposed an algorithm to produce coherent probabilistic forecasts and applied it to UK electricity smart meter data. Another study was carried out by Jeon et al. (2018) recently where they propose a novel method for probabilistic forecast reconciliation based on cross-validation which in particularly applied to the temporal hierarchies. Further, Gamakumara et al. (2018) define the coherent probabilistic forecasts, and forecast reconciliation proving an geometrical intuition to the problem. In the following sections we discuss these in detail.

4.1 Coherent probabilistic forecasts

Let us start with the definition given by Taieb et al. (2017) for coherent probabilistic forecasts.

Definition 4. (*As adapted from Taieb et al. (2017)*)
Let $\mathbf{a}_t \in \mathbb{R}^{n-m}$ *is a vector consisting series with different levels of aggregation, where* $\mathbf{a}_t = \mathbf{A}\mathbf{b}_t$. $\mathbf{A}$ *is a matrix containing the rows in* $\mathbf{S}$ *that are correspond to the aggregate levels. In that way* $\mathbf{S}$ *can be written as* $\mathbf{S}' = [\mathbf{A}' \; \mathbf{I}_m]$ *where* $\mathbf{I}_m$ *is the* m^{th} *order identity matrix. Thus we can write* $\mathbf{y}_t = (\mathbf{a}_t \; \mathbf{b}_t)'$. *Further let* $\hat{\mathbf{F}}_{\mathbf{b},T+h}$ *be h-step ahead joint predictive distribution for* $\mathbf{b}_{T+h}$, *where* $\hat{F}_{i,T+h}(y|\mathbf{y}_1,...,\mathbf{y}_T) = \mathbb{P}(y_{i,T+h} \leq y|\mathbf{b}_1,...,\mathbf{b}_T)$. *Let* $\hat{F}_{a_j,T+h}$ *be the predictive distribution for* $a_{j,T+h}$ *and* s_j *be the* j^{th} *row vector of* $\mathbf{S}$ *for* $j = 1,...,n-m$. *Then the joint predictive distribution of the hierarchy is said to be probabilistically coherent if* $a_{j,T+h} \overset{d}{=} s_j \mathbf{b}_{T+h}$ *where* $\overset{d}{=}$ *denote the equality in distributions.*

The above definition is based on the convolution of probability distributions. That is the predictive distributions of a hierarchy is said to be coherent, if the convolution

of forecast distributions of disaggregate series is equal to the forecast distribution of the corresponding aggregate series. Following this definition, they introduced a new algorithm to produce coherent probabilistic forecasts. They first generate, a sample from the bottom level predictive distribution, and then aggregated to obtain coherent probabilistic forecasts of the upper levels of the hierarchy. Initially they use MinT algorithm to reconcile the means of the bottom level forecast distributions, and then a copula-based approach is employed to model the dependency structure of the hierarchy. Resulting multi-dimensional distribution is used to generate the empirical forecast distributions for all bottom-level series. Thus, while Ben Taieb et al. (2017) provide coherent probabilistic forecasts, they do no forecast reconciliation of the distributions. Because they do not use all the information of from the hierarchy when producing coherent forecasts. In that sense, their approach is analogous to bottom-up point forecasting rather than forecast reconciliation.

Gamakumara et al. (2018) gives a different definition for coherent forecasts based on a geometrical interpretation which can be also extended to the probabilistic forecast reconciliation. This definition is given below.

Definition 5. (*As adapted from Gamakumara et al. (2018)*)

Suppose $(\mathbb{R}^m, \mathscr{F}_{\mathbb{R}^m}, \nu)$ *is a probability triple, where* $\mathscr{F}_{\mathbb{R}^m}$ *is the usual Borel* σ*-algebra on* $\mathbb{R}^m$*. Further let* $\breve{\nu}$ *be a probability measure on* $\mathfrak{s}$ *with* σ*-algebra* $\mathscr{F}_{\mathfrak{s}}$*. Here* $\mathscr{F}_{\mathfrak{s}}$ *is a collection of sets* $s(\mathscr{B})$*, where* $s(\mathscr{B})$ *denotes the image of the set* $\mathscr{B} \in \mathscr{F}_{\mathbb{R}^m}$ *under the mapping* $s(.)$*. Then measure* $\breve{\nu}$ *is coherent if it has the property*

$$\breve{\nu}(s(\mathscr{B})) = \nu(\mathscr{B}) \quad \forall \mathscr{B} \in \mathscr{F}_{\mathbb{R}^m},$$

Suppose any set $\mathscr{B} \in \mathscr{F}_{\mathbb{R}^m}$ is mapped to the coherent subspace $\mathfrak{s}$ through the mapping $s(.)$ as depicted in figure 1. Then the probability measure of $\mathscr{B}$ in $(\mathbb{R}^m, \mathscr{F}_{\mathbb{R}^m})$

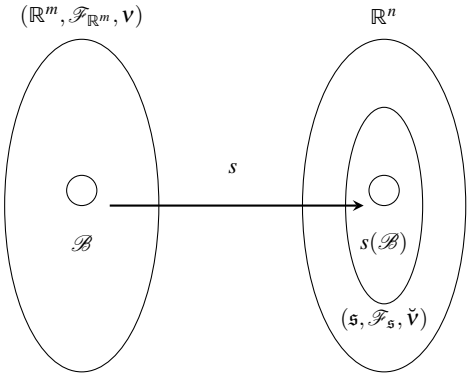

Fig. 1 Any set $\mathscr{B} \in \mathbb{R}^m$ will be mapped to $\mathfrak{s}$ through the mapping $s(.)$

is equivalent to that of $s(\mathscr{B})$ in $(\mathfrak{s}, \mathscr{F}_{\mathfrak{s}})$. Thus it follows from the definition, if the uncertainty of $\mathbf{y}_{t+h|t}$ is characterised by the probability triple $(\mathfrak{s}, \mathscr{F}_{\mathfrak{s}}, \check{\nu})$, then the probabilistic forecasts at time $t+h$ is said to be coherent. In turn this implies that there is no density of any $\mathbf{y}_{t+h|t}$ in the null space of $\mathfrak{s}$.

Although both definitions are conceptually consistent, the latter provide a geometrical intuition which is extended to the probabilistic forecast reconciliation as discussed in the following section.

4.2 Probabilistic forecast reconciliation

Let $(\mathbb{R}^m, \mathscr{F}_{\mathbb{R}^m}, \nu)$ and $(\mathbb{R}^n, \mathscr{F}_{\mathbb{R}^n}, \hat{\nu})$ are the probability measures defined on $\mathbb{R}^m$ and $\mathbb{R}^n$ spaces respectively. Latter characterises the uncertainty of incoherent point forecast $\hat{\mathbf{y}}_{t+h|t}$. This contains all the information of the hierarchy by considering the data upto time t, which however does not satisfy the aggregate constraints of the hierarchy. Thus it can be also referred to as the incoherent probability triple. Further $g : \mathbb{R}^n \to \mathbb{R}^m$ is a linear function which maps any set of points in $\mathbb{R}^n$ to $\mathbb{R}^m$.

Definition 6. (*As adapted from Gamakumara et al. (2018)*)
The reconciled probability measure of $\hat{\nu}$ with respect to the mapping $g(.)$ is a probability measure $\tilde{\nu}$ on $\mathfrak{s}$ with σ-algebra $\mathscr{F}_{\mathfrak{s}}$ such that

$$\tilde{\nu}(s(\mathscr{B})) = \nu(\mathscr{B}) = \hat{\nu}(g^{-1}(\mathscr{B})) \qquad \forall \mathscr{B} \in \mathscr{F}_{\mathbb{R}^m}\,, \tag{28}$$

where $g^{-1}(\mathscr{B}) := \{\check{\mathbf{y}} \in \mathbb{R}^n : g(\check{\mathbf{y}}) \in \mathscr{B}\}$ is the pre-image of $\mathscr{B}$, that is the set of all points in $\mathbb{R}^n$ that $g(.)$ maps to a point in $\mathscr{B}$.

Recall that in point forecast reconciliation we start with a set of incoherent forecasts which were obtained by independently fitting models to all series in the hierarchy. Then these forecasts were projected to the coherent subspace $\mathfrak{s}$ through the linear function $g(.)$ followed by $s(.)$. Thus we have $\tilde{\mathbf{y}} = \mathbf{SG}\hat{\mathbf{y}}$. Analogous to this, the above definition for probabilistic forecast reconciliation implies that the probability measure of a set of incoherent forecasts is equal to the probability measure of same points after linearly projecting them to the coherent subspace.

As depicted in figure 2, a set of incoherent forecasts will be mapped to a set $\mathscr{B} \in \mathbb{R}^m$ through the mapping $g(.)$. Thus the probability measure of set $\mathscr{B}$ with respect to ν is same as that of $g^{-1}(\mathscr{B})$ with respect to $\hat{\nu}$. This is analogous to projecting the incoherent set of point forecasts to the bottom level series in the point case. Next this set will be again mapped to the coherent subspace $\mathfrak{s}$ through the mapping $s(.)$. Then the probability measure of $s(\mathscr{B})$ with respect to $\tilde{\nu}$ is equal to the probability measure of $\mathscr{B} \in \mathbb{R}^m$ with respect to ν. Therefore we have the expression given in (28).

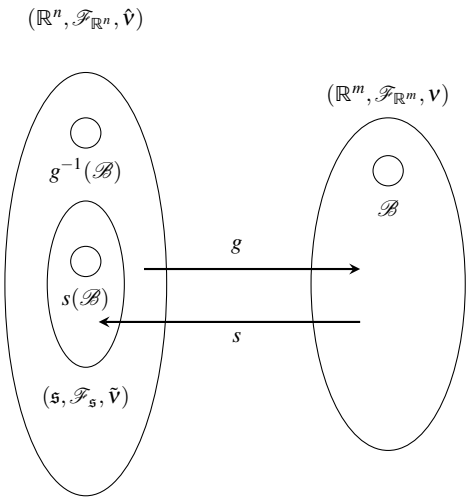

Fig. 2 Set of all points in $\mathbb{R}^n$ is mapped to a set $\mathscr{B} \in \mathbb{R}^m$ through the mapping $g(.)$. Then these will be again mapped to the coherent subspace $\mathfrak{s}$ through the mapping $s(.)$.

Let us now see how this definition can be used in reconciling probabilistic forecasts given the predictive densities. Recall that the point forecast reconciliation through projections is mainly follow three steps. Initially we change the coordinates of $\hat{\mathbf{y}}_{t+h}$ to $(\tilde{\mathbf{b}}'_{t+h}, \tilde{\mathbf{a}}'_{t+h})$ with respect to the basis $(\mathbf{S} \;\; \mathbf{R})$. Then we get $\tilde{\mathbf{b}}'_{t+h}$, the reconciled forecasts of bottom level series by eliminating the $\tilde{\mathbf{a}}'_{t+h}$ by setting $\tilde{\mathbf{a}}'_{t+h} = 0$. Finally we project these bottom level forecasts to the coherent subspace through $\tilde{\mathbf{y}}_{t+h} = \mathbf{S}\tilde{\mathbf{b}}_{t+h}$.

We can follow similar steps when reconciling the densities for a given hierarchy. Suppose standard notation for probability density function is given by $f(.)$. Then following equation (14) and the standard results for densities of transformed variables we have,

$$f(\hat{\mathbf{y}}_{t+h|t}) = f(\mathbf{S}\tilde{\mathbf{b}}_{t+h|t} + \mathbf{R}\tilde{\mathbf{a}}_{t+h|t})|(\mathbf{S}\;\mathbf{R})| \tag{29}$$

where $|.|$ denotes the determinant of a matrix.

The incoherent probability measure of the set $g^{-1}(\mathscr{B})$, i.e $\hat{\nu}(g^{-1}(\mathscr{B}))$ given in Definition 6 can be written as follows.

$$\hat{\nu}(g^{-1}(\mathscr{B})) = Pr(\hat{\mathbf{y}}_{t+h|t} \in g^{-1}(\mathscr{B})) = \int_{g^{-1}(\mathscr{B})} f(\hat{\mathbf{y}}_{t+h|t}) d\hat{\mathbf{y}}_{t+h|t} \tag{30}$$

Following equation (29) we have,

$$\hat{\nu}(g^{-1}(\mathscr{B})) = \int_{\mathscr{B}} \int f(\mathbf{S}\tilde{\mathbf{b}}_{t+h|t} + \mathbf{R}\tilde{\mathbf{a}}_{t+h|t})|(\mathbf{S}\;\mathbf{R})| d\tilde{\mathbf{b}} d\tilde{\mathbf{a}}. \tag{31}$$

This step corresponds to the change of coordinates in the point forecast reconciliation. Now to eliminate the coordinates of null space of coherent subspace, we simply marginalise over the null space. That is we integrate equation (31) with respect to $\tilde{\mathbf{a}}_{t+h}$. This step is analogous to equating $\tilde{\mathbf{a}}_{t+h}$ in the case of point reconciliation. Then we get the probability density corresponds to the reconciled bottom level series such that,

$$\nu(\mathscr{B}) = Pr(\tilde{\mathbf{b}}_{t+h|t} \in \mathscr{B}) = \int_{\mathscr{B}} f(\tilde{\mathbf{b}}_{t+h|t}) d\tilde{\mathbf{b}}_{t+h|t}. \tag{32}$$

In order to get the reconciled probability density of the whole hierarchy we simply follow the Definition 5. Thus we have,

$$\tilde{\nu}(s(\mathscr{B})) = Pr(\tilde{\mathbf{y}}_{t+h|t} \in s(\mathscr{B})) = \int_{s(\mathscr{B})} f(\tilde{\mathbf{y}}_{t+h|t}) d\tilde{\mathbf{y}}_{t+h|t}, \quad \tilde{\mathbf{y}}_{t+h|t} = \mathbf{S}\tilde{\mathbf{b}}_{t+h|t}. \tag{33}$$

The following subsection illustrates how this method can be used to reconcile an incoherent Gaussian forecast distribution.

4.2.1 Reconciliation of Gaussian forecast distributions

Suppose $\mathscr{N}(\hat{\mu}_{t+h}, \hat{\Sigma}_{t+h}) \overset{d}{\leftrightarrow} f(\hat{\mathbf{y}}_{t+h})$ is an incoherent forecast distribution at time $t+h$. Then from (29) it follows that

$$f(\hat{\mathbf{y}}_{t+h}) = f(\mathbf{S}\tilde{\mathbf{b}}_{t+h} + \mathbf{R}\tilde{\mathbf{t}}_{t+h}) \left|\mathbf{S} \;\; \mathbf{R}\right| = \frac{f(\mathbf{S}\tilde{\mathbf{b}}_{t+h} + \mathbf{R}\tilde{\mathbf{t}}_{t+h})}{\left|(\mathbf{S} \;\; \mathbf{R})^{-1}\right|}.$$

By substituting the Gaussian distribution function for $f(\hat{\mathbf{y}}_{t+h})$ we get,

$$\mathbf{f_B}(\cdot) = \frac{\exp\left\{-\frac{1}{2}(\mathbf{S}\tilde{\mathbf{b}}_{t+h} + \mathbf{R}\tilde{\mathbf{t}}_{t+h} - \hat{\mu}_{t+h})' \hat{\boldsymbol{\Sigma}}_{\mathbf{t+h}}^{-1} (\mathbf{S}\tilde{\mathbf{b}}_{t+h} + \mathbf{R}\tilde{\mathbf{t}}_{t+h} - \hat{\mu}_{t+h})\right\}}{(2\pi)^{\frac{n}{2}} \left|\hat{\Sigma}_{t+h}\right|^{\frac{1}{2}} \left|(\mathbf{S} \vdots \mathbf{R})^{-1}\right|}, \tag{34}$$

$$= \frac{\exp\left\{-\frac{1}{2}\left((\mathbf{S} \vdots \mathbf{R})\begin{pmatrix}\tilde{\mathbf{b}} \\ \mathbf{a}\end{pmatrix} - \hat{\mu}_{t+h}\right)' \hat{\boldsymbol{\Sigma}}_{\mathbf{t+h}}^{-1} \left((\mathbf{S} \vdots \mathbf{R})\begin{pmatrix}\tilde{\mathbf{b}} \\ \mathbf{a}\end{pmatrix} - \hat{\mu}_{t+h}\right)\right\}}{(2\pi)^{\frac{n}{2}} \left|\hat{\Sigma}_{t+h}\right|^{\frac{1}{2}} \left|(\mathbf{S} \vdots \mathbf{R})^{-1}\right|}, \tag{35}$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}} \left|\hat{\Sigma}_{t+h}\right|^{\frac{1}{2}} \left|(\mathbf{S} \vdots \mathbf{R})^{-1}\right|} \exp\left\{-\frac{1}{2}\left(\begin{pmatrix}\tilde{\mathbf{b}} \\ \mathbf{a}\end{pmatrix} - (\mathbf{S} \vdots \mathbf{R})^{-1}\hat{\mu}_{t+h}\right)'\right. \tag{36}$$

$$\left.\left[(\mathbf{S} \vdots \mathbf{R})\hat{\boldsymbol{\Sigma}}_{\mathbf{t+h}}(\mathbf{S} \vdots \mathbf{R})'\right]^{-1}\left(\begin{pmatrix}\tilde{\mathbf{b}} \\ \mathbf{a}\end{pmatrix} - (\mathbf{S} \vdots \mathbf{R})^{-1}\hat{\mu}_{t+h}\right)\right\} \tag{37}$$

Recall that

$$(\mathbf{S} \vdots \mathbf{R})^{-1} = \begin{pmatrix} (\mathbf{R}_\perp' \mathbf{S})^{-1} \mathbf{R}_\perp' \\ \cdots \\ (\mathbf{S}_\perp' \mathbf{R})^{-1} \mathbf{S}_\perp' \end{pmatrix} = \begin{pmatrix} \mathbf{P} \\ \mathbf{Q} \end{pmatrix},$$

where $\mathbf{P} = (\mathbf{R}_\perp' \mathbf{S})^{-1} \mathbf{R}_\perp'$ and $\mathbf{Q} = (\mathbf{S}_\perp' \mathbf{R})^{-1} \mathbf{S}_\perp'$. Then

$$\mathbf{f_B}(\cdot) = \frac{1}{(2\pi)^{\frac{n}{2}} \left| \hat{\Sigma}_{t+h} \right|^{\frac{1}{2}} \left| \begin{pmatrix} \mathbf{G} \\ \mathbf{H} \end{pmatrix} \right|} \exp\Big\{ -\frac{1}{2} \Big[\begin{pmatrix} \tilde{\mathbf{b}} \\ \mathbf{a} \end{pmatrix} - \begin{pmatrix} \mathbf{G} \\ \mathbf{H} \end{pmatrix} \hat{\mu}_{t+h} \Big]' \Big[\begin{pmatrix} \mathbf{G} \\ \mathbf{H} \end{pmatrix} \hat{\boldsymbol{\Sigma}}_{\mathbf{t+h}} \begin{pmatrix} \mathbf{G} \\ \mathbf{H} \end{pmatrix}' \Big]^{-1} \Big[\begin{pmatrix} \tilde{\mathbf{b}} \\ \mathbf{a} \end{pmatrix} - \begin{pmatrix} \mathbf{G} \\ \mathbf{H} \end{pmatrix} \hat{\mu}_{t+h} \Big] \Big\},$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}} \left| \begin{pmatrix} \mathbf{G} \\ \mathbf{H} \end{pmatrix} \hat{\boldsymbol{\Sigma}}_{\mathbf{t+h}} \begin{pmatrix} \mathbf{G} \\ \mathbf{H} \end{pmatrix}' \right|^{\frac{1}{2}}} \exp\Big\{ -\frac{1}{2} \begin{pmatrix} \tilde{\mathbf{b}}_{t+h} - \mathbf{P}\hat{\mu}_{t+h} \\ \tilde{\mathbf{t}}_{t+h} - \mathbf{Q}\hat{\mu}_{t+h} \end{pmatrix}' \Big[\begin{pmatrix} \mathbf{G} \\ \mathbf{H} \end{pmatrix} \hat{\boldsymbol{\Sigma}}_{\mathbf{t+h}} \begin{pmatrix} \mathbf{G} \\ \mathbf{H} \end{pmatrix}' \Big]^{-1} \begin{pmatrix} \tilde{\mathbf{b}}_{t+h} - \mathbf{P}\hat{\mu}_{t+h} \\ \tilde{\mathbf{t}}_{t+h} - \mathbf{Q}\hat{\mu}_{t+h} \end{pmatrix} \Big\}.$$

Since $\Big[\begin{pmatrix} \mathbf{G} \\ \mathbf{H} \end{pmatrix} \hat{\boldsymbol{\Sigma}}_{\mathbf{t+h}} \begin{pmatrix} \mathbf{G} \\ \mathbf{H} \end{pmatrix}' \Big] = \begin{pmatrix} \mathbf{P}\hat{\boldsymbol{\Sigma}}_{\mathbf{t+h}}\mathbf{P}' & \mathbf{P}\hat{\boldsymbol{\Sigma}}_{\mathbf{t+h}}\mathbf{Q}' \\ \mathbf{Q}\hat{\boldsymbol{\Sigma}}_{\mathbf{t+h}}\mathbf{P}' & \mathbf{Q}\hat{\boldsymbol{\Sigma}}_{\mathbf{t+h}}\mathbf{Q}' \end{pmatrix}$ we have

$$\mathbf{f_B}(\cdot) = \frac{1}{(2\pi)^{\frac{n}{2}} \left| \begin{pmatrix} \mathbf{P}\hat{\boldsymbol{\Sigma}}_{\mathbf{t+h}}\mathbf{P}' & \mathbf{P}\hat{\boldsymbol{\Sigma}}_{\mathbf{t+h}}\mathbf{Q}' \\ \mathbf{Q}\hat{\boldsymbol{\Sigma}}_{\mathbf{t+h}}\mathbf{P}' & \mathbf{Q}\hat{\boldsymbol{\Sigma}}_{\mathbf{t+h}}\mathbf{Q}' \end{pmatrix} \right|^{\frac{1}{2}}} \exp\Big\{ -\frac{1}{2} \begin{pmatrix} \tilde{\mathbf{b}}_{t+h} - \mathbf{P}\hat{\mu}_{t+h} \\ \tilde{\mathbf{t}}_{t+h} - \mathbf{Q}\hat{\mu}_{t+h} \end{pmatrix}' \begin{pmatrix} \mathbf{P}\hat{\boldsymbol{\Sigma}}_{\mathbf{t+h}}\mathbf{P}' & \mathbf{P}\hat{\boldsymbol{\Sigma}}_{\mathbf{t+h}}\mathbf{Q}' \\ \mathbf{Q}\hat{\boldsymbol{\Sigma}}_{\mathbf{t+h}}\mathbf{P}' & \mathbf{Q}\hat{\boldsymbol{\Sigma}}_{\mathbf{t+h}}\mathbf{Q}' \end{pmatrix}^{-1} \begin{pmatrix} \tilde{\mathbf{b}}_{t+h} - \mathbf{P}\hat{\mu}_{t+h} \\ \tilde{\mathbf{t}}_{t+h} - \mathbf{Q}\hat{\mu}_{t+h} \end{pmatrix} \Big\}.$$

This is the joint multivariate Gaussian distribution of $(\tilde{\mathbf{b}}_{t+h}' \vdots \tilde{\mathbf{t}}_{t+h}')'$. Then from (**??**) and the properties of the multivariate Gaussian distribution, it follows that

$$\tilde{\mathbf{f}}(\tilde{\mathbf{b}}_{t+h}) = \frac{1}{(2\pi)^{\frac{n}{2}} \left| \mathbf{P}\hat{\boldsymbol{\Sigma}}_{\mathbf{t+h}}\mathbf{P}' \right|^{\frac{1}{2}}} \exp\Big\{ -\frac{1}{2} (\tilde{\mathbf{b}}_{t+h} - \mathbf{P}\hat{\mu}_{t+h})' (\mathbf{P}\hat{\boldsymbol{\Sigma}}_{\mathbf{t+h}}\mathbf{P}')^{-1} (\tilde{\mathbf{b}}_{t+h} - \mathbf{P}\hat{\mu}_{t+h}) \Big\}. \tag{38}$$

Equation (38) implies $\tilde{\mathbf{b}}_{t+h} \sim \mathcal{N}(\mathbf{P}\hat{\mu}_{t+h}, \mathbf{P}\hat{\Sigma}_{t+h}\mathbf{P}')$ where $\mathbf{P} = (\mathbf{R}_\perp' \mathbf{S})^{-1} \mathbf{R}_\perp'$. Then from (**??**) it follows that

$$\tilde{\mathbf{f}}(\tilde{\mathbf{y}}_{t+h}) = \mathbf{S} \circ \tilde{\mathbf{f}}(\tilde{\mathbf{b}}_{t+h}) = \tilde{\mathbf{f}}(\mathbf{S}\tilde{\mathbf{b}}_{t+h}). \tag{39}$$

Therefore, the reconciled Gaussian forecast distribution of the whole hierarchy is $\mathcal{N}(\mathbf{SP}\hat{\mu}_{t+h},\ \mathbf{SP}\hat{\Sigma}_{t+h}\mathbf{P}'\mathbf{S}')$.

Reconciliation of parametric distributions - Gaussian case
Non-parametric bootstrap approach
Write about this article Jeon et al. (2018)

5 Hierarchical forecasts evaluation

5.1 Point forecast evaluation

Include the possible methods in a table.

5.2 Probabilistic forecast evaluation

Rephrase the following three paragraphs

The necessary final step in hierarchical forecasting is to make sure that our forecast distributions are accurate enough to predict the uncertain future. In general, forecasters prefer to maximize the sharpness of the predictive distribution subject to the calibration Gneiting & Katzfuss (2014). Therefore the probabilistic forecasts should be evaluated with respect to these two properties.

Calibration refers to the statistical compatibility between probabilistic forecasts and realizations. In other words, random draws from a perfectly calibrated predictive distribution should be equivalent to the realizations. On the other hand, sharpness refers to the spread or the concentration of prediction distributions and it is a property of forecasts only. The more concentrated the predictive distributions, the sharper the forecasts are Gneiting et al. (2008). However, independently assessing the calibration and sharpness will not help to properly evaluate the probabilistic forecasts. Therefore to assess these properties simultaneously, we use scoring rules.

Even though the log score can be used evaluate simulated forecast densities with large samples **?**, it is more convenient to use if it is reasonable to assume a parametric forecast density for the hierarchy. However, the "degenerecy" of coherent forecast densities would be problematic when using log scores.

5.2.1 Scoring rules

Scoring rules are summary measures obtained based on the relationship between predictive distribution and the realizations. In some studies, researchers take the scoring rules to be positively oriented which they would wish to maximize Gneiting & Raftery (2007). However, scoring rules were also defined to be negatively oriented which forecasters wish to minimize Gneiting & Katzfuss (2014). We consider these

negatively oriented scoring rules to evaluate probabilistic forecasts in hierarchical time series.

Let $\check{\mathbf{Y}}$ and $\mathbf{Y}$ be a n-dimensional random vectors from the predictive distribution $\mathbf{F}$ and the true distribution G. Further let $\mathbf{y}$ be a n-dimensional realization. Then the scoring rule is a numerical value $S(\check{\mathbf{Y}},\mathbf{y})$ assign to each pair $(\check{\mathbf{Y}},\mathbf{y})$ and the proper scoring rule is defined as,

$$\mathrm{e}_{\mathbf{G}}[S(\mathbf{Y},\mathbf{y})] \leq \mathrm{e}_{\mathbf{G}}[S(\check{\mathbf{Y}},\mathbf{y})], \tag{40}$$

where $\mathrm{e}_{\mathbf{G}}[S(\mathbf{Y},\mathbf{y})]$ is the expected score under the true distribution $\mathbf{G}$ Gneiting et al. (2008), Gneiting & Katzfuss (2014).

Following are few scoring rules which have been widely used to assess probabilistic forecasts in literature.

5.2.2 Univariate scoring rules

Univariate log score

Continuous Ranked Probability Score (CRPS)

CRPS is defined in terms of predictive cumulative distribution function (CDF) for evaluating univariate probabilistic forecasts and given as,

$$CRPS(F,y) = E_F\left|X-y\right| - \frac{1}{2}E_F\left|X-X'\right|, \tag{41}$$

where $y \in \mathbb{R}$ and X and X' are independent random variables from the forecast distribution F with finite first moment Gneiting & Raftery (2007). It reduces to the absolute error if F is a point forecast and therefore CRPS is meaningful to use to compare probabilistic forecasts and point forecasts Gneiting & Katzfuss (2014).

5.2.3 Multivariate scoring rules

Multivariate log score

Energy score

The multivariate generalization of CRPS is the energy score proposed by Gneiting et al. (2008) and is given by,

$$ES(\mathbf{F},\mathbf{y}) = E_{\mathbf{F}}||\mathbf{X}-\mathbf{y}|| - \frac{1}{2}E_{\mathbf{F}}||\mathbf{X}-\mathbf{X}'||, \tag{42}$$

where $\mathbf{y} \in \mathbb{R}^{\mathrm{d}}$ is the vector of realizations, $\mathbf{X}$ and $\mathbf{X}'$ are independent d dimension random vectors from the multivariate forecast distribution $\mathbf{F}$ and $||.||$ denotes the Euclidean norm. In many cases it is difficult to find the closed form expression for $ES(\mathbf{F},\mathbf{y})$ and hence the Monte Carlo methods will be employed. Gneiting et al. (2008) has further given the Monte Carlo approximation to the equation (42) as,

$$\hat{ES}(\mathbf{F},\mathbf{y}) = \frac{1}{k}\sum_{j=1}^{k}||\mathbf{x}_i - \mathbf{y}|| - \frac{1}{2(k-1)}\sum_{j=1}^{k}||\mathbf{x}_j - \mathbf{x}_{j+1}||, \tag{43}$$

where $\mathbf{x}_1, ..., \mathbf{x}_k$ is a simple random sample of size k (possibly large) from the predictive density $\mathbf{F}$.

However, Pinson (2013) has shown that energy score has a very low discrimination ability for incorrectly specified covariances even though it discriminates well in misspecified means.

Variogram score

6 Empirical study

Use the standard `equation` environment to typeset your equations, e.g.

$$a \times b = c \,, \tag{44}$$

however, for multiline equations we recommend to use the `eqnarray` environment[1].

$$\begin{aligned} a \times b &= c \\ \mathbf{a} \cdot \mathbf{b} &= \mathbf{c} \end{aligned} \tag{45}$$

> Please do not use quotation marks when quoting texts! Simply use the `quotation` environment – it will automatically render Springer's preferred layout.

Run-in Heading Boldface Version Use the LaTeX automatism for all your cross-references and citations as has already been described in Sect. **??**.

Run-in Heading Italic Version Use the LaTeX automatism for all your cross-references and citations as has already been described in Sect. **??**.

Table 1 Please write your table caption here

Classes	Subclass	Length	Action Mechanism
Translation	mRNA[a]	22 (19–25)	Translation repression, mRNA cleavage
Translation	mRNA cleavage	21	mRNA cleavage
Translation	mRNA	21–22	mRNA cleavage
Translation	mRNA	24–26	Histone and DNA Modification

[a] Table foot note (with superscript)

Acknowledgements If you want to include acknowledgments of assistance and the like at the end of an individual chapter please use the `acknowledgement` environment – it will automatically render Springer's preferred layout.

Appendix

When placed at the end of a chapter or contribution (as opposed to at the end of the book), the numbering of tables, figures, and equations in the appendix section continues on from that in the main text. Hence please *do not* use the `appendix` command when writing an appendix at the end of your chapter or contribution. If there is only one the appendix is designated "Appendix", or "Appendix 1", or "Appendix 2", etc. if there is more than one.

[1] In physics texts please activate the class option `vecphys` to depict your vectors in ***boldface-italic*** type - as is customary for a wide range of physical subjects

$$a \times b = c \tag{46}$$

References

Athanasopoulos, G., Ahmed, R. A. & Hyndman, R. J. (2009), 'Hierarchical forecasts for Australian domestic tourism', *International Journal of Forecasting* **25**(1), 146–166.

Ben Taieb, S., Huser, R., Hyndman, R. J. & Genton, M. G. (2017), 'Forecasting uncertainty in electricity smart meter data by boosting additive quantile regression', *IEEE Transactions on Smart Grid* **7**(5), 2448–2455.

Dunn, D. M., Williams, W. H. & Dechaine, T. L. (1976), 'Aggregate Versus Subaggregate Models in Local Area Forecasting', *Journal of American Statistical Association* **71**(353), 68–71.

Gamakumara, P., Panagiotelis, A., Athanasopoulos, G. & Hyndman, R. J. (2018), Probabilisitic Forecasts in Hierarchical Time Series.

Gel, Y., Raftery, A. E. & Gneiting, T. (2004), 'Calibrated Probabilistic Mesoscale Weather Field Forecasting', *Journal of the American Statistical Association* **99**(July), 575–583.

Gneiting, T. & Katzfuss, M. (2014), 'Probabilistic Forecasting', *Annual Review of Statistics and Its Application* **1**, 125–151.

Gneiting, T. & Raftery, A. E. (2005), 'Weather_forecasting_with_ensem.PDF', *Science* **310.5746**, 248–249.

Gneiting, T. & Raftery, A. E. (2007), 'Strictly Proper Scoring Rules, Prediction, and Estimation', *Journal of the American Statistical Association* **102**(477), 359–378.

Gneiting, T., Raftery, A. E., Westveld, A. H. & Goldman, T. (2005), 'Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation', *Monthly Weather Review* **133**(5), 1098–1118.

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L. & Johnson, N. A. (2008), 'Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds', *Test* **17**(2), 211–235.

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

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., Lee, A. J. & Wang, E. (2016), 'Fast computation of reconciled forecasts for hierarchical and grouped time series', *Computational Statistics and Data Analysis* **97**, 16–32.
URL: *http://dx.doi.org/10.1016/j.csda.2015.11.007*

Jeon, J., Panagiotelis, A. & Petropoulos, F. (2018), Reconciliation of probabilistic forecasts with an application to wind power.

McSharry, P. E., Bouwman, S. & Bloemhof, G. (2005), 'Probabilistic forecasts of the magnitude and timing of peak electricity demand', *IEEE Transactions on Power Systems* **20**(2), 1166–1172.

Mirčetić, D., Nikoličić, S., Stojanović, D. & Maslarić, M. (2017), 'Modified top down approach for hierarchical forecasting in a beverage supply chain', *Transportation Research Procedia* **22**, 193–202.

Pinson, P. (2013), 'Wind Energy: Forecasting Challenges for Its Operational Management', *Statistical Science* **28**(4), 564–585.
URL: *http://projecteuclid.org/euclid.ss/1386078879*

Pinson, P., Madsen, H., Papaefthymiou, G. & Klöckl, B. (2009), 'From Probabilistic Forecasts to Wind Power Production', *Wind Energy* **12**(1), 51–62.

Schäfer, J. & Strimmer, K. (2005), 'A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics', *Statistical Applications in Genetics and Molecular Biology* **4**(1).
URL: *https://www.degruyter.com/view/j/sagmb.2005.4.issue-1/sagmb.2005.4.1.1175/sagmb.2005.4.1.1175.xml*

Taieb, S. B., Taylor, J. W. & Hyndman, R. J. (2017), 'Hierarchical Probabilistic Forecasting of Electricity Demand with Smart Meter Data', pp. 1–30.

van Erven, T. & Cugliari, J. (2014), *Game-Theoretically Optimal reconciliation of contemporaneous hierarchical time series forecasts*.

Wickramasuriya, S. L., Athanasopoulos, G. & Hyndman, R. J. (2018), 'Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization', *Journal of the American Statistical Association* **1459**, 1–45.
URL: *https://www.tandfonline.com/doi/full/10.1080/01621459.2018.1448825*