ISSN 1440-771X

MONASH
BUSINESS
SCHOOL

Department of Econometrics and Business Statistics

http://business.monash.edu/econometrics-and-business-statistics/research/publications

Probabilistic Forecasts in Hierarchical Time Series

Puwasala Gamakumara
Anastasios Panagiotelis
George Athanasopoulos
Rob J Hyndman

March 2018

Working Paper ??/??

Probabilistic Forecasts in Hierarchical Time Series

Puwasala Gamakumara
Department of Econometrics and Business Statistics,
Monash University,
VIC 3800, Australia.
Email: Puwasala.Gamakumara@monash.edu

Anastasios Panagiotelis
Department of Econometrics and Business Statistics,
Monash University,
VIC 3800, Australia.
Email: Anastasios.Panagiotelis@monash.edu

George Athanasopoulos
Department of Econometrics and Business Statistics,
Monash University,
VIC 3800, Australia.
Email: George.Athanasopoulos@monash.edu

Rob J Hyndman
Department of Econometrics and Business Statistics,
Monash University,
VIC 3800, Australia.
Email: Rob.Hyndman@monash.edu

6 March 2018

JEL classification: ??

Probabilistic Forecasts in Hierarchical Time Series

Abstract

TBC

1 Introduction

Many research applications involve a large collection of time series, some of which are aggregates of others. These are called hierarchical time series. For example, electricity demand of a country can be disaggregated along a geographical hierarchy: the electricity demand of the whole country can be divided into the demand of states, cities, and households.

When forecasting such time series, it is important to have "coherent" forecasts across the hierarchy: aggregates of the forecasts at lower levels should be equal to the forecasts at the upper levels of aggregation. In other words, sums of forecasts should be equal to the forecasts of the sums.

The traditional approaches to produce coherent point forecasts are the bottom-up, top-down and middle-out methods. In the bottom-up approach, forecasts of the lowest level are first generated and they are simply aggregated to forecast upper levels of the hierarchy (Dunn, Williams, and Dechaine, 1976). In contrast, the top-down approach involves forecasting the most aggregated series first and then disaggregating these forecasts down the hierarchy based on the corresponding proportions of observed data (Gross and Sohl, 1990). Many studies have discussed the relative advantages and disadvantages of bottom-up and top-down methods, and situations in which each would provide reliable forecasts (Schwarzkopf, Tersine, and Morris, 1988; Kahn, 1998; Lapide, 1998; Fliedner, 2001). 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.

It is apparent that these three approaches use only part of the information available when producing coherent forecasts. This might result in inaccurate forecasts. For example, if the bottom level series are highly volatile or noisy, and hence challenging to forecast, then the resulting forecasts from the bottom-up approach are likely to be inaccurate.

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. The approach was further improved by Wickramasuriya, Athanasopoulos, and Hyndman (2017) who proposed the "MinT" algorithm to obtain optimally reconciled point forecasts by minimizing the mean squared coherent forecast errors.

Traditional bottom-up, top-down and middle-out forecasting methods are not strictly reconciliation methods since they use only a part of the information from the hierarchy to produce coherent 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 and Katzfuss, 2014). For example, McSharry, Bouwman, and Bloemhof (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, Raftery, and Gneiting (2004), Gneiting et al. (2005) and Gneiting and Raftery (2005) for various weather variables.

Although there is a rich and growing literature on producing coherent point forecasts of hierarchical time series, little attention has been given to coherent probabilistic forecasts. The only relevant paper we are aware of is Ben Taieb et al. (2017), who recently proposed an algorithm to produce coherent probabilistic forecasts and applied it to UK electricity smart meter data. In their approach, a sample from the bottom level predictive distribution is first generated, and then aggregated to obtain coherent probabilistic forecasts of the upper levels of the hierarchy. Hence this method is a bottom-up approach. They propose to first use the 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. The resulting multi-dimensional distribution is used to generating 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. In that sense, their approach is analogous to bottom-up point forecasting rather than forecast reconciliation.

After introducing our notation in Section 2, we define what is meant by probabilistic forecast reconciliation for hierarchical time series in Section 3. First, we provide a new definition for coherency of point forecasts, and the reconciliation of a set of incoherent point forecasts, using concepts related to vector spaces and measure theory. Based on these, we provide a rigorous definition for probabilistic forecast reconciliation, and how we can reconcile the incoherent forecast densities in practice.

Further, due to the aggregation structure of the hierarchy, the probability distribution is degenerate and hence the forecast distribution should also be degenerate. In Section 4, we discuss in detail how this degeneracy will be taken care of in probabilistic forecast reconciliation, and in Section 5 we consider the evaluation of probabilistic hierarchical forecasts.

Some theoretical results on probabilistic forecast reconciliation in the Gaussian framework are given in Section 6, including a simulation study to show the importance of reconciliation in the probabilistic framework.

We conclude with some thoughts on extensions and limitations in Section 7.

2 Notation

Our notation mostly follows that introduced in Wickramasuriya, Athanasopoulos, and Hyndman (2017). Suppose $\boldsymbol{y}_t \in \mathbb{R}^n$ comprising all observations of the whole hierarchy at time t and $\boldsymbol{b}_t \in \mathbb{R}^m$ comprising only the bottom level observations at time t. Then due to the aggregation nature of the hierarchy we have,

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

where $\boldsymbol{S}$ is a $n \times m$ constant matrix whose columns span the linear subspace for which all constraints hold.

To understand the notations clearly, consider the hierarchy given in Figure 1.

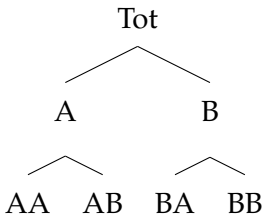

Figure 1: *Two level hierarchical diagram*

In any hierarchy, the most aggregated level is termed as level 0, the second most aggregated level is termed as level 1 and so on. 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 $\boldsymbol{y}_t = [y_{Tot,t}, y_{A,t}, y_{B,t}, y_{C,t}, y_{AA,t}, y_{AB,t}, y_{BA,t}, y_{BB,t}]'$,
$\boldsymbol{b}_t = [y_{AA,t}, y_{AB,t}, y_{BA,t}, y_{BB,t}]'$, $m = 4$, $n = 7$, and

$$\boldsymbol{S} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ & & I_4 & \end{pmatrix},$$

where I_4 is a 4-dimension identity matrix.

3 Coherent forecasts

The main purpose of this section is to provide formal definitions for coherent forecasts. We first give an alternative definition for coherent point forecasts using the properties of vector spaces which is then followed by the definition of coherent probabilistic forecasts.

Definition 3.1 (Coherent subspace) *Let $\mathbb{C}^m$ is a m-dimensional subspace of $\mathbb{R}^n$, $\mathbb{C}^m < \mathbb{R}^n$. $\mathbb{C}^m$ is said to be a coherent space, if it is spanned by the columns of $\boldsymbol{S}$.*

It is worth to notice that the coherent space $\mathbb{C}^m$ is equivalent to the column space of $\boldsymbol{S}$ which we denote by $\mathscr{C}(\boldsymbol{S})$. Further the space orthogonal to $\mathbb{C}^m$ is denoted by $\mathbb{N}^{n-m}$ and it is equivalent to the null space of $\boldsymbol{S}$.

Definition 3.2 (Coherent Point Forecasts) *Suppose $\breve{\boldsymbol{y}}_{t+h} \in \mathbb{R}^n$ consists point forecasts of each series in the hierarchy at time $t+h$. $\breve{\boldsymbol{y}}_{t+h}$ is said to be Coherent if $\breve{\boldsymbol{y}}_{t+h} \in \mathbb{C}$.*

Definition 3.3 (Coherent Probabilistic Forecasts) *Let* $(\mathbb{R}^m, \mathscr{F}^m, \nu^m)$ *be a probability triple where* $\mathscr{F}^m$ *is sigma algebra on* $\mathbb{R}^m$*. Then,* $(\mathbb{C}, \mathscr{F}_S, \breve{\nu})$ *is said to be coherent probability measure space iff,*

$$\breve{\nu}(\boldsymbol{S}(\boldsymbol{A})) = \nu^m(\boldsymbol{A}) \quad \forall \quad \boldsymbol{A} \in \mathscr{F}^m,$$

where $\boldsymbol{S}(\boldsymbol{A})$ *denotes the image of subset* $\boldsymbol{A}$ *under* $\boldsymbol{S}$*.*

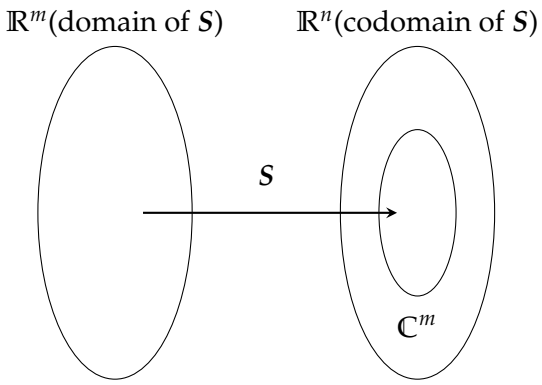

Figure 2: *Any set* $\boldsymbol{A} \in \mathbb{R}^m$ *will be mapped to the* $\mathbb{C}^m$ *through the mapping* $\boldsymbol{S}$

Definition (3.3) implies the probability measure on $\mathbb{C}^m$ is equivalent to the probability measure on $(\mathbb{R}^m, \mathscr{F}^m)$. Hence, there is no density anywhere outside the linear subspace $\mathbb{C}^m$. That is, a *Coherent probability density forecast* is any density $f(\breve{\boldsymbol{y}}_{t+h})$ such that $f(\breve{\boldsymbol{y}}_{t+h}) = 0$ for all $\breve{\boldsymbol{y}}_{t+h} \in \mathbb{N}^{n-m}$. Following example will help to understand these definitions more clearly.

Example 1

Consider a simple hierarchy with two bottom level series A and B that add up to the top level series Tot. Suppose the forecasts at time $t+h$ of these series are given by $\breve{\boldsymbol{y}}_{t+h} = [\breve{y}_{Tot,t+h}, \breve{y}_{A,t+h}, \breve{y}_{B,t+h}]$ and $\breve{\boldsymbol{y}}_{t+h} \in \mathbb{R}^3$. Due to the aggregation constraint of the hierarchy we have $\breve{y}_{Tot,t+h} = \breve{y}_{A,t+h} + \breve{y}_{B,t+h}$. This implies, even though $\breve{\boldsymbol{y}}_{t+h}$ is in $\mathbb{R}^3$, the points actually lie in $\mathbb{C}^2$, which is a two dimensional subspace within that $\mathbb{R}^3$ space. Therefore, for any $\breve{\boldsymbol{y}}_{t+h} \in \mathbb{N}$ have a zero probability. I.e. $f(\breve{\boldsymbol{y}}_{t+h}) = 0$ for any $\breve{\boldsymbol{y}}_{t+h} \in \mathbb{N}$.

For a particular coherent subspace $\mathbb{C}^m$, there exist several distinct basis vectors. For example, in the smallest hierarchy considered above, $\left\{\begin{pmatrix}1 & 1 & 0\end{pmatrix}', \begin{pmatrix}1 & 0 & 1\end{pmatrix}'\right\}$, $\left\{\begin{pmatrix}1 & 0 & 1\end{pmatrix}', \begin{pmatrix}0 & 1 & -1\end{pmatrix}'\right\}$ and the singular value decomposition of these two are some

alternative basis vectors that spans the same $\mathbb{C}^m$. Given a basis for $\mathbb{C}^m$, every series of the hierarchy can be linearly determined as a linear combination of that basis vectors. We referred 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, $(\boldsymbol{A}, \boldsymbol{B})$ and $(\boldsymbol{Tot}, \boldsymbol{A})$ are the basis series corresponds to the basis vectors $\left\{\begin{pmatrix}1 & 1 & 0\end{pmatrix}', \begin{pmatrix}1 & 0 & 1\end{pmatrix}'\right\}$ and $\left\{\begin{pmatrix}1 & 0 & 1\end{pmatrix}', \begin{pmatrix}0 & 1 & -1\end{pmatrix}'\right\}$ respectively. Thus it is clear that the set of bottom level series is a basis series that corresponds to the column vectors of $\boldsymbol{S}$.

An important thing to notice here is, since the basis are not unique for a give coherent subspace, the definition (3.3) is not unique and one can redefine the coherent probabilistic forecasts with respect to any basis. However, we stick to the definition (3.3) and consider the basis defined by the columns of $\boldsymbol{S}$ in what follows.

It is also worth to mention that the definition (3.2) and (3.3) facilitate extension to the forecast reconciliation which we talk about in the next section. In contrast to our definition, Ben Taieb et al. (2017) defines the coherent probabilistic forecasts in terms of convolution. According to their definition, if the forecasts are coherent, then the convolution of forecast distributions of disaggregate series is same as the forecast distribution of the corresponding aggregate series.

4 Forecast reconciliation

Previous studies on point forecast literature have shown that the reconciliation provides better coherent forecasts than the conventional bottom-up and top-down methods (Hyndman et al., 2011; Erven and Cugliari, 2014; Wickramasuriya, Athanasopoulos, and Hyndman, 2017). However, not extended in the context of probabilistic forecasts in published literature. Thus our main focus is to rigorously define the probabilistic forecast reconciliation and discuss how that definition can be used in practice.

4.1 Point forecast reconciliation

Initially we define point forecast reconciliation as a groundwork to the probabilistic forecast reconciliation.

Definition 4.1 *Let $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^m$ and $\hat{\boldsymbol{y}}_{t+h} \in \mathbb{R}^n$ be any set of incoherent forecasts at time $t+h$. Then $\tilde{\boldsymbol{b}}_{t+h}$ is said to be reconciled bottom level forecasts if*

$$\tilde{\boldsymbol{b}}_{t+h} = \boldsymbol{g}(\hat{\boldsymbol{y}}_{t+h}),$$

where $\boldsymbol{g}(\hat{\boldsymbol{y}}_{t+h})$ *is the image of* $\hat{\boldsymbol{y}}_{t+h}$ *under* $\boldsymbol{g}$ *on* $\mathbb{R}^m$*. The reconciled forecasts for the whole hierarchy is then given by* $\tilde{\boldsymbol{y}}_{t+h} = \boldsymbol{S} \circ \boldsymbol{g}(\hat{\boldsymbol{y}}_{t+h})$ *such that* $\tilde{\boldsymbol{y}}_{t+h} \in \mathbb{C}^m$*, where* $\boldsymbol{S} \circ \boldsymbol{g}(.)$ *is a projection of* $\boldsymbol{g}(.)$ *onto the* $\mathbb{C}^m$*.*

Importance of definition (4.1) is that it facilitates for both linear and non-linear reconciliation. In other words, if $\boldsymbol{g}$ is a non-linear function, then the reconciliation of $\hat{\boldsymbol{y}}_{t+h}$ will be non-linear. On the other hand, if $\boldsymbol{g}$ is a linear function, then $\boldsymbol{S} \circ \boldsymbol{g}(.)$ will linearly projects incoherent point forecasts onto $\mathbb{C}^m$. Latter was the main focus in previous studies in hierarchical point forecasting, where $\boldsymbol{g}$ is considered as a $m \times n$ matrix $\boldsymbol{P}$ and thus $\tilde{\boldsymbol{y}}_{t+h} = \boldsymbol{SP}\hat{\boldsymbol{y}}_{t+h}$.

In the following content we provide an alternative explanation for linear reconciliation based on definition (4.1). Let $\boldsymbol{R} \in \mathbb{R}^{n \times (n-m)}$ consists the columns that spans $\mathbb{N}^{n-m}$ which is orthogonal to $\mathbb{C}^m$. Note that $\boldsymbol{R}$ is also not unique and one example is a matrix whose columns represent the aggregation constraints for a given hierarchy. Then for the hierarchy in example 1,

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

Further let $\{\boldsymbol{s}_1, ..., \boldsymbol{s}_m\}$ and $\{\boldsymbol{r}_1, ..., \boldsymbol{r}_{n-m}\}$ denote the columns of $\boldsymbol{S}$ and $\boldsymbol{R}$ respectively. Then $\boldsymbol{B} = \{\boldsymbol{s}_1, ..., \boldsymbol{s}_m, \boldsymbol{r}_1, ..., \boldsymbol{r}_{n-m}\}$ is a basis for $\mathbb{R}^n$. Now, using the insights of definition (4.1), 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 $\hat{\boldsymbol{y}}_{t+h} \in \mathbb{R}^n$, first we find the coordinates of $\hat{\boldsymbol{y}}_{t+h}$ with respect to the basis $\boldsymbol{B}$. Let $\begin{pmatrix} \tilde{\boldsymbol{b}}'_{t+h} & \vdots & \tilde{\boldsymbol{t}}'_{t+h} \end{pmatrix}'$ denote these coordinates. Note that $\tilde{\boldsymbol{b}}_{t+h}$ is a basis series which is really the reconciled bottom level series that corresponds to the coefficients of linear combination of the basis $\{\boldsymbol{s}_1, ..., \boldsymbol{s}_m\}$. Similarly, $\tilde{\boldsymbol{t}}_{t+h}$ is another basis series corresponds to the coefficients of linear combination of the basis $\{\boldsymbol{r}_1, ..., \boldsymbol{r}_{n-m}\}$. Then from basic properties of linear algebra it follows that,

$$\begin{pmatrix} \boldsymbol{S} & \vdots & \boldsymbol{R} \end{pmatrix} \begin{pmatrix} \tilde{\boldsymbol{b}}'_{t+h} & \vdots & \tilde{\boldsymbol{t}}'_{t+h} \end{pmatrix}' = \hat{\boldsymbol{y}}_{t+h},$$

$$\hat{\boldsymbol{y}}_{t+h} = \boldsymbol{S}\tilde{\boldsymbol{b}}_{t+h} + \boldsymbol{R}\tilde{\boldsymbol{t}}_{t+h},$$

and

$$\begin{pmatrix} \tilde{\boldsymbol{b}}'_{t+h} & \vdots & \tilde{\boldsymbol{t}}'_{t+h} \end{pmatrix}' = \begin{pmatrix} \boldsymbol{S} & \vdots & \boldsymbol{R} \end{pmatrix}^{-1} \hat{\boldsymbol{y}}_{t+h}. \tag{1}$$

In order to find $\begin{pmatrix} \boldsymbol{S} & \vdots & \boldsymbol{R} \end{pmatrix}^{-1}$, let $\boldsymbol{S}_\perp$ and $\boldsymbol{R}_\perp$ be the orthogonal complements of $\boldsymbol{S}$ and $\boldsymbol{R}$ respectively. Then $\begin{pmatrix} \boldsymbol{S} & \vdots & \boldsymbol{R} \end{pmatrix}^{-1}$ is given by,

$$\begin{pmatrix} \boldsymbol{S} & \vdots & \boldsymbol{R} \end{pmatrix}^{-1} = \begin{pmatrix} (\boldsymbol{R}'_\perp \boldsymbol{S})^{-1} \boldsymbol{R}'_\perp \\ \cdots \\ (\boldsymbol{S}'_\perp \boldsymbol{R})^{-1} \boldsymbol{S}'_\perp \end{pmatrix}.$$

Thus we have,

$$\begin{pmatrix} \tilde{\boldsymbol{b}}_{t+h} \\ \cdots \\ \tilde{\boldsymbol{t}}_{t+h} \end{pmatrix} = \begin{pmatrix} (\boldsymbol{R}'_\perp \boldsymbol{S})^{-1} \boldsymbol{R}'_\perp \\ \cdots \\ (\boldsymbol{S}'_\perp \boldsymbol{R})^{-1} \boldsymbol{S}'_\perp \end{pmatrix} \hat{\boldsymbol{y}}_{t+h}. \tag{2}$$

From (2) it follows that,

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

Step 2: Obtaining reconciled point forecasts for the whole hierarchy

This step directly follows by the definition for coherent forecasts. That is, to obtain reconciled point forecasts for the entire hierarchy, we map $\tilde{\boldsymbol{b}}_{t+h} \in \mathbb{R}^n$ to the $\mathbb{C}^m$ through $\boldsymbol{S}$. Thus we have,

$$\tilde{\boldsymbol{y}}_{t+h} = \boldsymbol{S}(\boldsymbol{R}'_\perp \boldsymbol{S})^{-1} \boldsymbol{R}'_\perp \hat{\boldsymbol{y}}_{t+h}, \quad \tilde{\boldsymbol{y}}_{t+h} \in \mathbb{C}^m < \mathbb{R}^n.$$

Finding a suitable $\boldsymbol{R}_\perp$ with respect to a certain loss function will result optimally reconciled point forecasts of the hierarchy. Notice that, if $\boldsymbol{P} = (\boldsymbol{R}'_\perp \boldsymbol{S})^{-1} \boldsymbol{R}'_\perp$, then the definition for linear reconciliation of point forecasts in previous studies coincides with our explanation.

Further in our context, we need to find $\boldsymbol{R}_\perp$ such that $\boldsymbol{R}'_\perp \boldsymbol{S}$ is invertible. i.e, $(\boldsymbol{R}'_\perp \boldsymbol{S})^{-1} \boldsymbol{R}'_\perp \boldsymbol{S} = \boldsymbol{I}$. This condition coincides with the unbiased condition $\boldsymbol{SPS} = \boldsymbol{S}$ proposed by Hyndman et al. (2011).

In their study, Hyndman et al. (2011) proposed to choose,

$$\tilde{\boldsymbol{b}}^{OLS}_{t+h} = (\boldsymbol{S}'\boldsymbol{S})^{-1} \boldsymbol{S}' \hat{\boldsymbol{y}}_{t+h},$$

where in this context, $\boldsymbol{R}_{\perp}' = \boldsymbol{S}'$. Thus the reconciled point forecasts for the entire hierarchy is given by,

$$\tilde{\boldsymbol{y}}_{t+h}^{OLS} = \boldsymbol{S}(\boldsymbol{S}'\boldsymbol{S})^{-1}\boldsymbol{S}'\hat{\boldsymbol{y}}_{t+h}.$$

They referred this to as OLS solution and the loss function they considered is equivalent to the euclidean norm between $\hat{\boldsymbol{y}}_{t+h}$ and $\tilde{\boldsymbol{y}}_{t+h}$, i.e. $< \hat{\boldsymbol{y}}_{T+h}, \tilde{\boldsymbol{y}}_{T+h} >$.

According to a recent study by Wickramasuriya, Athanasopoulos, and Hyndman (2017), choosing $\boldsymbol{R}_{\perp}' = \boldsymbol{S}'\boldsymbol{W}_h^{-1}$ will minimize the trace of mean squared reconciled forecast errors under the property of unbiasedness where, $\boldsymbol{W}_h^{-1}$ is the variance of the incoherent forecast errors. This will result,

$$\tilde{\boldsymbol{b}}_{t+h}^{MinT} = (\boldsymbol{S}'\boldsymbol{W}_h^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}_h^{-1}\hat{\boldsymbol{y}}_{t+h},$$

and thus,

$$\tilde{\boldsymbol{y}}_{t+h}^{MinT} = \boldsymbol{S}(\boldsymbol{S}'\boldsymbol{W}_h^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}_h^{-1}\hat{\boldsymbol{y}}_{t+h}.$$

They referred this to as MinT solution. It is also worth to notice that the loss function they considered is equivalent to the Mahalanobis distance between $\hat{\boldsymbol{y}}_{T+h}$ and $\tilde{\boldsymbol{y}}_{T+h}$. i.e. $< \hat{\boldsymbol{y}}_{T+h}, \tilde{\boldsymbol{y}}_{T+h} >_{W_h}$.

4.2 Probabilistic forecast reconciliation

In terms of probabilistic forecasts, the reconciliation implies finding the probability measure of the coherent forecasts using the information of incoherent probabilistic forecast measure. A more formal definition is given below.

Definition 4.2 *Suppose* $(\mathbb{R}^n, \mathscr{F}^n, \hat{\nu})$ *be an incoherent probability triple and* $(\mathbb{R}^m, \mathscr{F}^m, \nu^m)$ *be a probability triple defined on* $\mathbb{R}^m$. *Let* $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^m$. *Then the probability measure on reconciled bottom levels is such that,*

$$\nu^m(\boldsymbol{A}) = \hat{\nu}(\boldsymbol{g}^{-1}(\boldsymbol{A})), \quad \forall \quad \boldsymbol{A} \in \mathscr{F}^m.$$

Further the reconciled probability measure of the whole hierarchy is given by,

$$\tilde{\nu}(\boldsymbol{S}(\boldsymbol{A})) = \hat{\nu}(\boldsymbol{g}^{-1}(\boldsymbol{A})), \quad \forall \quad \boldsymbol{A} \in \mathscr{F}^m,$$

where $\boldsymbol{S} : \mathbb{R}^m \to \mathbb{C}^m$ *and* $\tilde{\nu}(.)$ *is the probability measure on the measure space* $(\mathbb{C}^m, \mathscr{F}_S)$.

Since the above definition seems not to be straightforward in reconciling incoherent forecasts, the following content explains how this can be used in practice to obtain reconciled probabilistic forecasts for hierarchical time series.

Recall that $\hat{\boldsymbol{y}}_{t+h}$ is a set of incoherent point forecasts and the coordinates of that with respect to the basis $\boldsymbol{B}$ is given by (1). Suppose $\hat{f}(.)$ is the probability density of $\hat{\boldsymbol{y}}_{t+h}$. Our goal is to reconcile $\hat{f}(.)$ such that the density lives on the $\mathbb{C}^m$. In order to obtain this reconciled density, we need to project $\hat{f}(\hat{\boldsymbol{y}}_{t+h})$ onto the $\mathbb{C}^m$ along the direction of $\mathbb{N}^{n-m}$.

Let the density of coordinates of $\hat{\boldsymbol{y}}_{t+h}$ with respect to basis $\boldsymbol{B}$ is denoted by $f_{\boldsymbol{B}}(.)$. Then it follows from (1) and the facts on density of transformed variables,

$$f_{\boldsymbol{B}}(\tilde{\boldsymbol{b}}_{t+h}, \tilde{\boldsymbol{t}}_{t+h}) = \hat{f}(\boldsymbol{S}\tilde{\boldsymbol{b}}_{t+h} + \boldsymbol{R}\tilde{\boldsymbol{t}}_{t+h}) \quad \left| \begin{array}{ccc} \boldsymbol{S} & \vdots & \boldsymbol{R} \end{array} \right|, \tag{3}$$

where $|.|$ denote the determinant of a matrix. Now that we have the density of $\begin{pmatrix} \tilde{\boldsymbol{b}}'_{t+h} & \vdots & \tilde{\boldsymbol{t}}'_{t+h} \end{pmatrix}'$, the marginal density of $\tilde{\boldsymbol{b}}_{t+h}$ can be obtained by integrating (3) over the range of $\tilde{\boldsymbol{t}}_{t+h}$. This will result the reconciled density of the bottom level series $\tilde{\boldsymbol{b}}_{t+h}$. i.e.,

$$\tilde{f}(\tilde{\boldsymbol{b}}_{t+h}) = \int_{lim(\tilde{\boldsymbol{t}}_{t+h})} \hat{f}(\boldsymbol{S}\tilde{\boldsymbol{b}}_{t+h} + \boldsymbol{R}\tilde{\boldsymbol{t}}_{t+h}) \quad \left| \begin{array}{ccc} \boldsymbol{S} & \vdots & \boldsymbol{R} \end{array} \right| \quad d\tilde{\boldsymbol{t}}_{t+h}. \tag{4}$$

Finally to get the reconciled density of the whole hierarchy, we simply follow the definition (3.3) and have,

$$\tilde{f}(\tilde{\boldsymbol{y}}_{t+h}) = \boldsymbol{S} \circ \tilde{f}(\tilde{\boldsymbol{b}}_{t+h}). \tag{5}$$

This final step will transform every point in the density $\tilde{f}(\tilde{\boldsymbol{b}}_{t+h})$ to the $\mathbb{C}^m < \mathbb{R}^n$. Following example illustrates how this method can be used to reconcile an incoherent Gaussian forecast distribution.

Example 2

Suppose $\mathcal{N}(\hat{\boldsymbol{\mu}}_{t+h}, \hat{\boldsymbol{\Sigma}}_{t+h}) \overset{d}{\leftrightarrow} \hat{f}(\hat{\boldsymbol{y}}_{t+h})$ is an incoherent forecast distribution at time $t+h$. Then from (3) it follows,

$$f_{\boldsymbol{B}}(\tilde{\boldsymbol{b}}_{t+h}, \tilde{\boldsymbol{t}}_{t+h}) = \hat{f}(\boldsymbol{S}\tilde{\boldsymbol{b}}_{t+h} + \boldsymbol{R}\tilde{\boldsymbol{t}}_{t+h}) \quad \left| \begin{array}{ccc} \boldsymbol{S} & \vdots & \boldsymbol{R} \end{array} \right| = \frac{\hat{f}(\boldsymbol{S}\tilde{\boldsymbol{b}}_{t+h} + \boldsymbol{R}\tilde{\boldsymbol{t}}_{t+h})}{\left| \begin{array}{ccc} \boldsymbol{S} & \vdots & \boldsymbol{R} \end{array} \right|^{-1}}.$$

By substituting the Gaussian distribution function to $f_{\boldsymbol{B}}(.)$ we get,

$$
\begin{aligned}
f_{\boldsymbol{B}}(.) &= \frac{\exp\left\{-\frac{1}{2}(\boldsymbol{S}\tilde{\boldsymbol{b}}_{t+h}+\boldsymbol{R}\tilde{\boldsymbol{t}}_{t+h}-\hat{\boldsymbol{\mu}}_{t+h})'\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t+h}}^{-1}(\boldsymbol{S}\tilde{\boldsymbol{b}}_{t+h}+\boldsymbol{R}\tilde{\boldsymbol{t}}_{t+h}-\hat{\boldsymbol{\mu}}_{t+h})\right\}}{(2\pi)^{\frac{n}{2}}\left|\hat{\boldsymbol{\Sigma}}_{t+h}\right|^{\frac{1}{2}}\left|\begin{pmatrix}\boldsymbol{S} & \vdots & \boldsymbol{R}\end{pmatrix}^{-1}\right|}, \\
&= \frac{\exp\left\{-\frac{1}{2}\left(\begin{pmatrix}\boldsymbol{S} & \vdots & \boldsymbol{R}\end{pmatrix}\begin{pmatrix}\tilde{\boldsymbol{b}}_{t+h}\\ \tilde{\boldsymbol{t}}_{t+h}\end{pmatrix}-\hat{\boldsymbol{\mu}}_{t+h}\right)'\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t+h}}^{-1}\left(\begin{pmatrix}\boldsymbol{S} & \vdots & \boldsymbol{R}\end{pmatrix}\begin{pmatrix}\tilde{\boldsymbol{b}}_{t+h}\\ \tilde{\boldsymbol{t}}_{t+h}\end{pmatrix}-\hat{\boldsymbol{\mu}}_{t+h}\right)\right\}}{(2\pi)^{\frac{n}{2}}\left|\hat{\boldsymbol{\Sigma}}_{t+h}\right|^{\frac{1}{2}}\left|\begin{pmatrix}\boldsymbol{S} & \vdots & \boldsymbol{R}\end{pmatrix}^{-1}\right|}, \\
&= \frac{1}{(2\pi)^{\frac{n}{2}}\left|\hat{\boldsymbol{\Sigma}}_{t+h}\right|^{\frac{1}{2}}\left|\begin{pmatrix}\boldsymbol{S} & \vdots & \boldsymbol{R}\end{pmatrix}^{-1}\right|}\exp\Big\{-\frac{1}{2}\Big(\begin{pmatrix}\tilde{\boldsymbol{b}}_{t+h}\\ \tilde{\boldsymbol{t}}_{t+h}\end{pmatrix}-\begin{pmatrix}\boldsymbol{S} & \vdots & \boldsymbol{R}\end{pmatrix}^{-1}\hat{\boldsymbol{\mu}}_{t+h}\Big)' \\
&\qquad\Big[\begin{pmatrix}\boldsymbol{S} & \vdots & \boldsymbol{R}\end{pmatrix}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t+h}}\begin{pmatrix}\boldsymbol{S} & \vdots & \boldsymbol{R}\end{pmatrix}'\Big]^{-1}\Big(\begin{pmatrix}\tilde{\boldsymbol{b}}_{t+h}\\ \tilde{\boldsymbol{t}}_{t+h}\end{pmatrix}-\begin{pmatrix}\boldsymbol{S} & \vdots & \boldsymbol{R}\end{pmatrix}^{-1}\hat{\boldsymbol{\mu}}_{t+h}\Big)\Big\}.
\end{aligned}
$$

Recall that,

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

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

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

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

Since, $\left[\begin{pmatrix} P \\ Q \end{pmatrix} \hat{\Sigma}_{t+h} \begin{pmatrix} P \\ Q \end{pmatrix}' \right] = \begin{pmatrix} P\hat{\Sigma}_{t+h}P' & P\hat{\Sigma}_{t+h}Q' \\ Q\hat{\Sigma}_{t+h}P' & Q\hat{\Sigma}_{t+h}Q' \end{pmatrix}$ we have,

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

$f_B(.)$ gives the joint distribution of $\begin{pmatrix} \tilde{b}'_{t+h} & \vdots & \tilde{t}'_{t+h} \end{pmatrix}'$, which is a multivariate Gaussian distribution. Then from (4) and the properties of marginalization of multivariate Gaussian distribution it follows,

$$\tilde{f}(\tilde{b}_{t+h}) = \frac{1}{(2\pi)^{\frac{n}{2}} \left| P\hat{\Sigma}_{t+h}P' \right|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2}(\tilde{b}_{t+h} - P\hat{\mu}_{t+h})'(P\hat{\Sigma}_{t+h}P')^{-1}(\tilde{b}_{t+h} - P\hat{\mu}_{t+h}) \right\}. \quad (6)$$

Equation (6) implies $\tilde{b}_{t+h} \sim \mathcal{N}(P\hat{\mu}_{t+h}, P\hat{\Sigma}_{t+h}P')$ where $P = (R'_\perp S)^{-1}R'_\perp$. Then from (5) it follows that,

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

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

5 Evaluation of hierarchical probabilistic forecasts

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 and 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.

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 and Raftery, 2007). However, scoring rules were also defined to be negatively oriented which forecasters wish to minimize (Gneiting and Katzfuss, 2014). We consider these negatively oriented scoring rules to evaluate probabilistic forecasts in hierarchical time series.

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

$$\mathrm{E}_{\boldsymbol{G}}[S(\boldsymbol{Y}, \boldsymbol{y})] \leq \mathrm{E}_{\boldsymbol{G}}[S(\breve{\boldsymbol{Y}}, \boldsymbol{y})], \tag{7}$$

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

Table 1 summarizes few existing proper scoring rules.

Table 1: *Scoring rules to evaluate multivariate forecast densities. $\breve{\boldsymbol{y}}_{T+h}$ and $\breve{\boldsymbol{y}}^{*}_{T+h}$ be two independent random vectors from the coherent forecast distribution $\breve{\boldsymbol{F}}$ with the density function $\breve{\boldsymbol{f}}(.)$ at time $T+h$ and $\boldsymbol{y}_{T+h}$ is the vector of realizations. Further $\breve{Y}_{T+h,i}$ and $\breve{Y}_{T+h,j}$ are ith and jth components of the vector $\breve{\boldsymbol{Y}}_{T+h}$. Further the variogram score is given for order p where, w_{ij} are non-negative weights.*

Scoring rule	**Expression**	**Reference**
Log score	$\mathrm{LS}(\breve{\boldsymbol{F}}, \boldsymbol{y}_{T+h}) = -\log \breve{\boldsymbol{f}}(\boldsymbol{y}_{T+h})$	Gneiting and Raftery (2007)
Energy score	$\mathrm{eS}(\breve{\boldsymbol{Y}}_{T+h}, \boldsymbol{y}_{T+h}) = \mathrm{E}_{\breve{\boldsymbol{F}}}\lVert\breve{\boldsymbol{Y}}_{T+h} - \boldsymbol{y}_{T+h}\rVert^{\alpha} - \frac{1}{2}\mathrm{E}_{\breve{\boldsymbol{F}}}\lVert\breve{\boldsymbol{Y}}_{T+h} - \breve{\boldsymbol{Y}}^{*}_{T+h}\rVert^{\alpha}, \quad \alpha \in (0,2]$	Gneiting et al. (2008)
Variogram score	$\mathrm{VS}(\breve{\boldsymbol{F}}, \boldsymbol{y}_{T+h}) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\Big(\lvert y_{T+h,i} - y_{T+h,j}\rvert^{p} - \mathrm{E}_{\breve{\boldsymbol{F}}}\lvert\breve{Y}_{T+h,i} - \breve{Y}_{T+h,j}\rvert^{p}\Big)^{2}$	Scheuerer and Hamill (2015)

Even though the log score can be used evaluate simulated forecast densities with large samples (Jordan, Krüger, and Lerch, 2017), 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. We will discuss more on this in the next subsection.

In the energy score, for $\alpha = 2$, it can be easily shown that

$$\text{eS}(\check{\boldsymbol{Y}}_{\boldsymbol{T+h}}, \boldsymbol{y}_{\boldsymbol{T+h}}) = \|\boldsymbol{y}_{T+h} - \check{\boldsymbol{\mu}}_{T+h}\|^2, \tag{8}$$

where $\check{\boldsymbol{\mu}}_{T+h} = \text{E}_{\boldsymbol{F}}(\check{\boldsymbol{Y}}_{T+h})$. Therefore in the limiting case, the energy score only measures the accuracy of the forecast mean, but not the entire distribution. Further Pinson and Tastu (2013) argued that the Energy score given in Table 1 has a very low discrimination ability for incorrectly specified covariances, even though it discriminates the misspecified means well.

However, Scheuerer and Hamill (2015) have shown that the variogram score has a high discrimination ability of misspecified means, variance and correlation structure than the Energy score. Further they suggested the variogram score with $p = 0.5$ is more powerful.

For a possible finite sample of size B from the multivariate forecast density $\check{\boldsymbol{F}}$, the variogram score is defined as,

$$\text{VS}(\check{\boldsymbol{F}}, \boldsymbol{y}_{T+h}) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(|y_{T+h,i} - y_{T+h,j}|^p - \frac{1}{B}\sum_{k=1}^{B}|\check{Y}^k_{T+h,i} - \check{Y}^k_{T+h,j}|^p\right)^2.$$

5.1 Evaluating coherent forecast densities

As it was mentioned in the previous section, any coherent hierarchical forecast density is a degenerate density. To the best of our knowledge, there is no proper multivariate scoring rule in literature to evaluate degenerate densities. Further it can be easily seen that some of the existing scoring rules breakdown under the degeneracy. For example take the log score in the univariate case. Suppose the true density is degenerate at $x = 0$, i.e. $f(x) = \mathbb{1}\{x = 0\}$. Now consider two predictive densities $p_1(x)$ and $p_2(x)$. Let $p_1(x)$ is equivalent to the true density, i.e. $p_1(x) = \mathbb{1}\{x = 0\}$ and $p_2(x) =^d N(0, \sigma^2)$ with $\sigma^2 < (2\pi)^{-1}$. The expected log score of p_1 is:

$$\text{E}_f[S(f, f)] = \text{E}_f[S(p_1, f)] = -\log[p_1(x = 0)] = 0,$$

and that of p_2 is:

$$\mathrm{E}_f[S(p_2, f)] = -\log[p_2(x = 0)] < 0.$$

Therefore $S(f, f) > S(p_2, f)$ and hence there exist at least one forecast density which breaks the condition (7) for proper scoring rule. This implies log score cannot be used to evaluate the degenerate densities.

Thus it is necessary to have a rule of thumb to use these scoring rules in order to evaluate coherent forecast densities. First we should notice that, even though the coherent distribution of the entire hierarchy is degenerate, the density of the basis set of series is non-degenerate since these series are linearly independent. Further, if we can correctly specify the forecast distribution of these basis set of series, then we have almost obtained the correct forecast distribution of the whole hierarchy. Therefore, we propose to evaluate the predictive ability of only the basis set of series of the coherent forecast density by using any of the above discussed multivariate scoring rules. This will also avoid the impact of degeneracy for the scoring rules.

For example, since the bottom level series is a set of basis series for a given hierarchy, we can evaluate the predictive ability of the bottom level series of the coherent forecast distribution instead of evaluating the whole distribution. Further, if our purpose is to compare two coherent forecast densities, we can compare the forecast ability of only the bottom level forecast densities.

5.2 Comparison of coherent and incoherent forecast densities

It is also important to assess how the coherent or reconciled forecast densities improve the predictive ability compared to the incoherent forecasts. Clearly, we cannot use multivariate scoring rules, even for the basis set of series, since the coherent and incoherent forecast densities lie in two different matrix spaces.

However we could compare individual margins of the forecast density of the hierarchy using univariate proper scoring rules. Most widely used Continuous Ranked Probability Score (CRPS) would be helpful for this.

$$\mathrm{CRPS}(\check{F}_i, y_{T+h,i}) = \mathrm{E}_{\check{F}_i}|\check{Y}_{T+h,i} - y_{T+h,i}| - \frac{1}{2}\mathrm{E}_{\check{F}_i}|\check{Y}_{T+h,i} - \check{Y}^*_{T+h,i}|,$$

where $\check{Y}_{T+h,i}$ and $\check{Y}^*_{T+h,i}$ are two independent copies from the ith reconciled marginal forecast distribution $\tilde{F}_i$ of the hierarchy and $y_{T+h,i}$ is the ith realization from the true marginal distribution

G_i. We can also use univariate log scores for which we could assume a parametric forecast distribution.

6 Probabilistic forecast reconciliation in the Gaussian framework

Main purpose of this section is to establish the importance of reconciliation in probabilistic hierarchical forecasting. We narrow down the simulation setting for the Gaussian framework in this work. That is, suppose all the historical data in the hierarchy follows a multivariate Gaussian distribution, i.e. $\boldsymbol{y}_T \sim \mathcal{N}(\boldsymbol{\mu}_T, \boldsymbol{\Sigma}_T)$ where both $\boldsymbol{\mu}_T$ and $\boldsymbol{\Sigma}_T$ lives in $\mathbb{C}^m$ by nature of the hierarchical time series. We are interested in estimating the predictive Gaussian distribution of $\boldsymbol{Y}_{T+h}|\boldsymbol{\mathcal{I}}_T$ where $\boldsymbol{\mathcal{I}}_T = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_T\}$, which should also lives in $\mathbb{C}^m$.

Considering the individual series in the hierarchy, it is well known that the optimal point forecasts with respect to the minimal mean square error is given by the $\mathrm{E}[Y_{T+h,i} \mid y_{1,i}, \ldots, y_{T,i}]$, $i = 1, \ldots, n$. Suppose we independently fit time series models for each series in the hierarchy. Then the point forecasts from the estimated models, denoted by $\hat{Y}_{T+h,i}$ is unbiased and consistent estimator of $\mathrm{E}[Y_{T+h,i} \mid y_{1,i}, \ldots, y_{T,i}]$, given that the parameter estimates of the fitted models are unbiased and asymptotically consistent.

For example, suppose the data from ith series follows a ARMA(p, q) model. i.e.,

$$Y_{t,i} = \alpha_1 Y_{t-1,i} + \cdots + \alpha_p Y_{t-p,i} + \epsilon_t + \beta_1 \epsilon_{t-1,i} + \cdots + \beta_q \epsilon_{t-q,i},$$

where $\epsilon_t \sim \mathcal{NID}(0, \sigma_i^2)$. Then,

$$\mathrm{E}[Y_{T+h,i} \mid y_{1,i}, \ldots, y_{T,i}] = \alpha_1 Y_{T+h-1,i} + \cdots + \alpha_p Y_{T+h-p,i} + \beta_1 \epsilon_{T+h-1,i} + \cdots + \beta_q \epsilon_{T+h-q,i}.$$

Since $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)'$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)'$ are unknown in practice and thus estimated using the maximum likelihood method. Let $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ denote the maximum likelihood estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ respectively. Yao and Brockwell (2006) showed that $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ are asymptotically consistent estimators. Thus the point forecasts from this estimated model, $\hat{Y}_{T+h,i}$, will also be a consistent estimator for $\mathrm{E}[Y_{T+h,i} \mid y_{1,i}, \ldots, y_{T,i}]$. i.e.,

$$\hat{Y}_{T+h,i} \xrightarrow{p} \mathrm{E}[Y_{T+h,i} \mid y_{1,i}, \ldots, y_{T,i}] \quad \text{as} \quad T \to \infty. \tag{9}$$

Let $\hat{\boldsymbol{Y}}_{T+h} = (\hat{Y}_{T+h,1}, \ldots, \hat{Y}_{T+h,n})'$ and (9) holds for all $i = 1, \ldots, n$. Then from Slutsky's theorem it follows that,

$$\hat{\boldsymbol{Y}}_{T+h} \xrightarrow{p} \mathrm{E}[\boldsymbol{Y}_{T+h}|\boldsymbol{\mathcal{I}}_T] \quad \text{as} \quad T \to \infty. \tag{10}$$

Further let the forecast error due to $\hat{\boldsymbol{Y}}_{T+h}$ is given by,

$$\hat{\boldsymbol{e}}_{T+h} = \boldsymbol{Y}_{T+h} - \hat{\boldsymbol{Y}}_{T+h},$$

Now consider the variance of $\hat{\boldsymbol{e}}_{T+h}$,

$$\begin{aligned}
\mathrm{E}[(\boldsymbol{Y}_{T+h} - \hat{\boldsymbol{Y}}_{T+h})(\boldsymbol{Y}_{T+h} - \hat{\boldsymbol{Y}}_{T+h})'|\boldsymbol{\mathcal{I}}_T] =& \mathrm{E}[(\boldsymbol{Y}_{T+h} - \mathrm{E}(\boldsymbol{Y}_{T+h}|\boldsymbol{\mathcal{I}}_T) + \mathrm{E}(\boldsymbol{Y}_{T+h}|\boldsymbol{\mathcal{I}}_T) - \hat{\boldsymbol{Y}}_{T+h}) \\
& (\boldsymbol{Y}_{T+h} - \mathrm{E}(\boldsymbol{Y}_{T+h}|\boldsymbol{\mathcal{I}}_T) + \mathrm{E}(\boldsymbol{Y}_{T+h}|\boldsymbol{\mathcal{I}}_T) - \hat{\boldsymbol{Y}}_{T+h})'|\boldsymbol{\mathcal{I}}_T], \\
=& \mathrm{E}[(\boldsymbol{Y}_{T+h} - \mathrm{E}(\boldsymbol{Y}_{T+h}|\boldsymbol{\mathcal{I}}_T))(\boldsymbol{Y}_{T+h} - \mathrm{E}(\boldsymbol{Y}_{T+h}|\boldsymbol{\mathcal{I}}_T))'|\boldsymbol{\mathcal{I}}_T] \\
& + \mathrm{E}[\mathrm{E}(\boldsymbol{Y}_{T+h}|\boldsymbol{\mathcal{I}}_T) - \hat{\boldsymbol{Y}}_{T+h})(\mathrm{E}(\boldsymbol{Y}_{T+h}|\boldsymbol{\mathcal{I}}_T) - \hat{\boldsymbol{Y}}_{T+h})'|\boldsymbol{\mathcal{I}}_T] \\
& + \mathrm{E}[(\boldsymbol{Y}_{T+h} - \mathrm{E}(\boldsymbol{Y}_{T+h}|\boldsymbol{\mathcal{I}}_T))(\mathrm{E}(\boldsymbol{Y}_{T+h}|\boldsymbol{\mathcal{I}}_T) - \hat{\boldsymbol{Y}}_{T+h})'|\boldsymbol{\mathcal{I}}_T] \\
& + \mathrm{E}[\mathrm{E}(\boldsymbol{Y}_{T+h}|\boldsymbol{\mathcal{I}}_T) - \hat{\boldsymbol{Y}}_{T+h})(\boldsymbol{Y}_{T+h} - \mathrm{E}(\boldsymbol{Y}_{T+h}|\boldsymbol{\mathcal{I}}_T))'|\boldsymbol{\mathcal{I}}_T]
\end{aligned}$$

From (10) it immediately follows that,

$$\mathrm{E}[(\boldsymbol{Y}_{T+h} - \hat{\boldsymbol{Y}}_{T+h})(\boldsymbol{Y}_{T+h} - \hat{\boldsymbol{Y}}_{T+h})'|\boldsymbol{\mathcal{I}}_T] \xrightarrow{p} \mathrm{E}[(\boldsymbol{Y}_{T+h} - \mathrm{E}(\boldsymbol{Y}_{T+h}|\boldsymbol{\mathcal{I}}_T))(\boldsymbol{Y}_{T+h} - \mathrm{E}(\boldsymbol{Y}_{T+h}|\boldsymbol{\mathcal{I}}_T))'|\boldsymbol{\mathcal{I}}_T],$$

$$\boldsymbol{W}_{T+h} \xrightarrow{p} \mathrm{Var}(\boldsymbol{Y}_{T+h}|\boldsymbol{\mathcal{I}}_T) \quad \text{as} \quad T \to \infty,$$

where, $\mathrm{E}[(\boldsymbol{Y}_{T+h} - \hat{\boldsymbol{Y}}_{T+h})(\boldsymbol{Y}_{T+h} - \hat{\boldsymbol{Y}}_{T+h})'|\boldsymbol{\mathcal{I}}_T] = \boldsymbol{W}_{T+h}$.

It should be noted that, even though $\hat{\boldsymbol{Y}}_{T+h}$ and $\boldsymbol{W}_{T+h}$ are asymptotically consistent estimators for $\mathrm{E}(\boldsymbol{Y}_{T+h}|\boldsymbol{\mathcal{I}}_T)$ and $\mathrm{Var}(\boldsymbol{Y}_{T+h}|\boldsymbol{\mathcal{I}}_T)$ respectively, they are not coherent since they doesn't lie in the coherent subspace. Thus the Gaussian forecast distribution with mean $\hat{\boldsymbol{Y}}_{T+h}$ and variance $\boldsymbol{W}_{T+h}$ will be incoherent and we denote it by,

$$\widehat{\boldsymbol{Y}_{T+h,i}|\boldsymbol{\mathcal{I}}_T} \sim \mathcal{N}(\hat{\boldsymbol{Y}}_{T+h}, \boldsymbol{W}_{T+h}) \tag{11}$$

Since our primary objective is to find the coherent forecast density of the hierarchy, we need to reconciled (11). Recalling from example 2, the reconciled Gaussian predictive distribution is

Table 2: *Summarizing different estimates of $\boldsymbol{R}'_{\perp}$. For $n < T$, $\hat{\boldsymbol{W}}^{sam}_{T+1}$ is an unbiased and consistent estimator for $\boldsymbol{W}_{T+1}$. $\hat{\boldsymbol{W}}^{shr}_{T+1}$ is a shrinkage estimator which is much suitable for large dimensions. $\hat{\boldsymbol{W}}^{shr}_{T+1}$ was proposed by Schäfer and Strimmer (2005) and also used by Wickramasuriya, Athanasopoulos, and Hyndman (2017), where $\tau = \frac{\sum_{i\neq j} \hat{Var}(\hat{r}_{ij})}{\sum_{i\neq j} \hat{r}^2_{ij}}$, $\hat{r}_{ij}$ is the ijth element of sample correlation matrix. Further Diag($\boldsymbol{A}$) denote the diagonal matrix of $\boldsymbol{A}$*

Method	**Estimate of W_h**	**Estimate of $R'_{\perp}$**
OLS	$\boldsymbol{I}$	$\boldsymbol{S}'$
MinT(Sample)	$\hat{\boldsymbol{W}}^{sam}_{T+1}$	$\boldsymbol{S}'(\hat{\boldsymbol{W}}^{sam}_{T+1})^{-1}$
MinT(Shrink)	$\hat{\boldsymbol{W}}^{shr}_{T+1} = \tau\text{Diag}(\hat{\boldsymbol{W}}^{sam}_{T+1}) + (1-\tau)\hat{\boldsymbol{W}}^{sam}_{T+1}$	$\boldsymbol{S}'(\hat{\boldsymbol{W}}^{shr}_{T+1})^{-1}$
MinT(WLS)	$\hat{\boldsymbol{W}}^{wls}_{T+1} = \text{Diag}(\hat{\boldsymbol{W}}^{shr}_{T+1})$	$\boldsymbol{S}'(\hat{\boldsymbol{W}}^{wls}_{T+1})^{-1}$

then given by,

$$\widetilde{\boldsymbol{Y}_{T+h,i}|\boldsymbol{\mathcal{I}}_T} \sim \mathcal{N}(\boldsymbol{SP}\hat{\boldsymbol{Y}}_{T+h}, \boldsymbol{SP}\boldsymbol{W}_{T+h}\boldsymbol{P}'\boldsymbol{S}')$$

where, $\boldsymbol{P} = (\boldsymbol{R}'_{\perp}\boldsymbol{S})^{-1}\boldsymbol{R}'_{\perp}$.

Result 1: Choosing $\boldsymbol{R}'_{\perp} = \boldsymbol{S}'\boldsymbol{W}^{-1}_{T+h}$ will ensure at least the mean of the predictive Gaussian distribution is optimally reconciled with respect to the energy score.

Result 1 can be easily shown as follows. From (8) the energy score at the upper limit of α is given by, $\|\boldsymbol{y}_{T+h} - \boldsymbol{SP}\hat{\boldsymbol{y}}_{T+h}\|^2$. Then the expectation of energy score with respect to the true distribution is equivalent to the trace of mean squared forecast error, i.e.

$$\text{E}_G[eS(\tilde{\boldsymbol{Y}}_{\boldsymbol{T+h}}, \boldsymbol{y}_{\boldsymbol{T+h}})] = \text{Tr}\{\text{E}_{\boldsymbol{y}_{T+h}}[(\boldsymbol{Y}_{T+h} - \boldsymbol{SP}\hat{\boldsymbol{Y}}_{T+h})(\boldsymbol{Y}_{T+h} - \boldsymbol{SP}\hat{\boldsymbol{Y}}_{T+h})'|\mathcal{I}_T]\}.$$

From Theorem 1 of Wickramasuriya, Athanasopoulos, and Hyndman (2017) it immediately follows that $\boldsymbol{P} = (\boldsymbol{S}'\boldsymbol{W}^{-1}_{T+h}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}^{-1}_{T+h}$ minimizes the expected energy score constrained on the unbiasedness of reconciled forecasts. Thus we have $\boldsymbol{R}'_{\perp} = \boldsymbol{S}'\boldsymbol{W}^{-1}_{T+h}$.

It should be noted that $\boldsymbol{W}_{T+h}$ can be estimated in different methods which yields different estimation of $\boldsymbol{R}'_{\perp}$. Table 2 summarizes these methods.

It is worth mentioning that all these forecasting methods were well established in the context of point forecast reconciliation (Hyndman et al., 2011; Wickramasuriya, Athanasopoulos, and Hyndman, 2017; Hyndman, Lee, and Wang, 2016). However our attempt is to emphasis the use

of these reconciliation methods in the context of probabilistic forecasts at least in the Gaussian framework.

Simulation setup

We consider the hierarchy given in figure (1) for this simulation study. This hierarchy consists two aggregation levels with four bottom level series. Each bottom-level series will be generated first and add them up to obtain the data for respective upper-level series. Hierarchical time series in practice contain much noisier series in the bottom level than in aggregate series. In order to simulate this feature in the hierarchy, we refer to Wickramasuriya, Athanasopoulos, and Hyndman (2017) and the data generating process will be given as follows.

Suppose $\{w_{AA,t}, w_{AB,t}, w_{BA,t}, w_{BB,t}\}$ are generated from $ARIMA(p,d,q)$ processes where, (p,q) and d take integers from $\{1,2\}$ and $\{0,1\}$ respectively with equal probability. Further, the contemporaneous errors $\{\epsilon_{AA,t}, \epsilon_{AB,t}, \epsilon_{BA,t}, \epsilon_{BB,t}\} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. The parameters for AR and MA components will be randomly and uniformly generated from $[0.3, 0.5]$ and $[0.3, 0.7]$ respectively. Then the bottom level series $\{y_{AA,t}, y_{AB,t}, y_{BA,t}, y_{BB,t}\}$ will be obtained as:

$$
\begin{aligned}
y_{AA,t} &= w_{AA,t} + u_t - 0.5v_t, \\
y_{AB,t} &= w_{AB,t} - u_t - 0.5v_t, \\
y_{BA,t} &= w_{BA,t} + u_t + 0.5v_t, \\
y_{BB,t} &= w_{BB,t} - u_t + 0.5v_t,
\end{aligned}
$$

where $u_t \sim N(0, \sigma_u^2)$ and $v_t \sim N(0, \sigma_v^2)$.

To obtain the aggregate series at level 1, we add their respective bottom level series such as:

$$
\begin{aligned}
y_{A,t} &= w_{AA,t} + w_{AB,t} - v_t, \\
y_{B,t} &= w_{BA,t} + w_{BB,t} + v_t,
\end{aligned}
$$

and the total series will be obtained as:

$$
y_{Tot,t} = w_{AA,t} + w_{AB,t} + w_{BA,t} + w_{BB,t}.
$$

To get less noisier aggregate series than disaggregate series, we choose $\mathbf{\Sigma}, \sigma_u^2$ and σ_v^2 such that,

$$\text{Var}(\epsilon_{AA,t} + \epsilon_{AB,t} + \epsilon_{BA,t} + \epsilon_{BB,t}) \leq \text{Var}(\epsilon_{AA,t} + \epsilon_{AB,t} - v_t) \leq \text{Var}(\epsilon_{AA,t} + u_t - 0.5v_t),$$

$$\boldsymbol{l}_1\mathbf{\Sigma}\boldsymbol{l}_1' \leq \boldsymbol{l}_2\mathbf{\Sigma}\boldsymbol{l}_2' + \sigma_v^2 \leq \boldsymbol{l}_3\mathbf{\Sigma}\boldsymbol{l}_3' + \sigma_u^2 + \frac{1}{4}\sigma_v^2,$$

where $\boldsymbol{l}_1 = \begin{pmatrix} 1 & 1 & 1 & 1 \end{pmatrix}, \boldsymbol{l}_2 = \begin{pmatrix} 1 & 1 & 0 & 0 \end{pmatrix}$ and $\boldsymbol{l}_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \end{pmatrix}$.

This follows,

$$\boldsymbol{l}_1\mathbf{\Sigma}\boldsymbol{l}_1' - \boldsymbol{l}_2\mathbf{\Sigma}\boldsymbol{l}_2' \leq \sigma_v^2 \leq \frac{4}{3}(\sigma_u^2 + \boldsymbol{l}_3\mathbf{\Sigma}\boldsymbol{l}_3' - \boldsymbol{l}_2\mathbf{\Sigma}\boldsymbol{l}_2').$$

Thus we choose, $\mathbf{\Sigma} = \begin{pmatrix} 5.0 & 3.1 & 0.6 & 0.4 \\ 3.1 & 4.0 & 0.9 & 1.4 \\ 0.6 & 0.9 & 2.0 & 1.8 \\ 0.4 & 1.4 & 1.8 & 3.0 \end{pmatrix}$, $\sigma_u^2 = 19$ and $\sigma_u^2 = 18$ in our simulation setting.

As such we generate data for the hierarchy with sample size $T = 501$. Then univariate $ARIMA$ models were fitted for each series independently using the first 500 observations and obtain 1-step ahead base (incoherent) forecasts. We use *forecast* package in **R**-software Hyndman (2017) for model fitting and forecasting. Further, different estimates of $\boldsymbol{W}_{T+1}$ and the corresponding $\boldsymbol{R}_\perp'$ were obtained as summarized in Table 2. This process was then replicated using 1000 different data sets from the same data generating process.

To assess the predictive performance of different forecasting methods, we use scoring rules as discussed in Section 5. In addition to that we use Skill score (Gneiting and Raftery, 2007) for any comparison. For a given forecasting method, evaluated by a particular scoring rule $S(.)$, the skill score will be calculated as follows,

$$Ss[S_B(.)] = \frac{S_B(\boldsymbol{Y}, \boldsymbol{y})^{\text{ref}} - S_B(\check{\boldsymbol{Y}}, \boldsymbol{y})}{S_B(\boldsymbol{Y}, \boldsymbol{y})^{\text{ref}}} \times 100\%,$$

where $S_B(.)$ is average score over B samples and $S_B(\boldsymbol{Y}, \boldsymbol{y})^{\text{ref}}$ is the average score of the reference forecasting methods. Thus $Ss[S_B(.)]$ gives the percentage improvement of the preferred forecasting method relative to the reference method. Any negative value of $Ss[S_B(.)]$ indicate that the method we compared is poor than the reference method, whereas any positive value indicates that method is superior to the reference method.

As it was mentioned before we wish to establish the importance of reconciliation methods from this simulation study. In particular, we compare different reconciliation methods over the

Table 3: *Comparison of incoherent forecasts using bottom level series. The "Skill score" columns give the percentage skill score with reference to the bottom up forecasting method. A positive entry in these columns shows the percentage increase of score for different reconciliation methods with relative to the bottom up method.*

Forecasting	Energy score		Log score		Variogram score	
method	Mean score	Skill score	Mean score	Skill score	Mean score	Skill score
MinT(Shrink)	7.47	10.11	11.34	6.44	3.05	4.69
MinT(Sample)	7.47	10.11	11.33	6.52	3.05	4.69
MinT(WLS)	7.91	4.81	12.64	−4.29	3.23	−0.94
OLS	10.14	−22.02	135.13	−1014.93	4.60	−43.75
Bottom up	8.31		12.12		3.20	

Table 4: *Comparison of incoherent vs coherent forecasts for the aggregate series using Skill score. "Incoherent" row represent the average score for incoherent forecasts. Each entry above this row represent the percentage skill score with reference to the incoherent forecasts. A positive(negative) entry shows the percentage increase(decrease) of score for different forecasting methods with relative to incoherent forecasts.*

Forecasting	Total		Series - A		Series - B	
method	CRPS	LogS	CRPS	LogS	CRPS	LogS
MinT(Shrink)	1.12	0.34	10.07	2.93	5.41	1.52
MinT(Sample)	1.12	0.34	10.07	2.93	5.41	1.52
MinT(WLS)	−2.61	−2.02	5.28	−4.40	2.70	−4.24
OLS	−38.06	−698.99	−24.70	−1368.33	−24.86	−1159.09
Bottom up	−89.55	−21.83	−8.87	−2.35	−9.46	−2.73
Incoherent	*2.68*	*2.97*	*4.17*	*3.41*	*3.70*	*3.30*

conventional bottom-up method and also evaluate the predictive ability of coherent forecasts over incoherent forecasts. For the former comparison, we use bottom level probabilistic forecasts and calculate the percentage skill score based on energy score, log score and variogram score for each reconciliation method with reference to the bottom up method (presented in Table 3). For the latter comparison, we use percentage skill score based on CRPS and univariate log score for coherent probabilistic forecasts of each individual series with reference to incoherent forecasts (presented in Tables 4 and 5).

It is clearly evident from the results in Table 3 that the multivariate reconciled forecasts for the bottom level series from MinT(Shrink) and MinT(Sample) outperform the bottom-up forecasts. Further, these two methods produce probabilistic forecasts with best predictive ability in comparison to incoherent forecasts (from Tables 4 and 5). Moreover, it turns out that OLS and bottom-up methods produce the worst forecasts.

Table 5: *Comparison of incoherent vs coherent forecasts for the individual bottom level series using Skill score.*

Forecasting	Series - AA		Series - AB		Series - BA		Series - BB	
method	CRPS	LogS	CRPS	LogS	CRPS	LogS	CRPS	LogS
MinT(Shrink)	8.71	2.71	10.57	3.04	5.95	1.86	7.91	2.46
MinT(Sample)	8.71	2.71	10.57	3.04	5.95	1.86	8.19	2.46
MinT(WLS)	5.54	0.30	5.96	0.30	2.43	−0.62	5.08	0.62
OLS	−22.43	−931.63	−22.49	−886.32	−26.01	−834.67	−23.45	−812.92
Incoherent	*3.79*	*3.32*	*3.69*	*3.29*	*3.46*	*3.23*	*3.54*	*3.25*

7 Conclusions

Although the problem of hierarchical point forecasts is well studied in the literature, there is a lack of attention in the context of probabilistic forecasts. Thus we attempted to fill this gap in the literature by providing substantial theoretical background to the problem. We initially provided rigorous definitions for the coherent point and probabilistic forecasts using the principles of measure theory. Due to the aggregation nature of hierarchy, the probability density is a degenerate density. Thus the forecast distribution that we opt to find should also lie in a lower dimensional subspace of $\mathbb{R}^n$.

As it was well established that the reconciliation outperforms other conventional point forecasting methods in the hierarchical literature, we proposed to use reconciliation in probabilistic framework to obtain coherent degenerate densities. We provided a distinct definition for density forecast reconciliation and how it can be used to reconcile incoherent densities in practice.

Assuming a multivariate Gaussian distribution for the hierarchy, we showed how to obtain reconciled Gaussian forecast densities, utilizing available information in the hierarchy. An extensive Monte Carlo simulation study further showed that the MinT reconciliation method (Wickramasuriya, Athanasopoulos, and Hyndman, 2017) is useful in producing improved coherent probabilistic forecasts at least in the Gaussian framework.

References

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

Dunn, DM, Williams, WH, and Dechaine, TL (1976). Aggregate Versus Subaggregate Models in Local Area Forecasting. *Journal of American Statistical Association* **71**(353), 68–71.

Erven, T van and Cugliari, J (2014). *Game-Theoretically Optimal reconciliation of contemporaneous hierarchical time series forecasts*. Ed. by A Antoniadis, X Brossat, and J Poggi, pp. 297–317.

Fliedner, G (2001). Hierarchical forecasting: issues and use guidelines. *Industrial Management & Data Systems* **101**(1), 5–12.

Gel, Y, Raftery, AE, and Gneiting, T (2004). Calibrated Probabilistic Mesoscale Weather Field Forecasting. *Journal of the American Statistical Association* **99**(July), 575–583.

Gneiting, T and Katzfuss, M (2014). Probabilistic Forecasting. *Annual Review of Statistics and Its Application* **1**, 125–151.

Gneiting, T and Raftery, AE (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association* **102**(477), 359–378.

Gneiting, T, Raftery, AE, Westveld, AH, and Goldman, T (2005). Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation. *Monthly Weather Review* **133**(5), 1098–1118.

Gneiting, T and Raftery, AE (2005). Weather_forecasting_with_ensem.PDF. *Science* **310.5746**, 248–249.

Gneiting, T, Stanberry, LI, Grimit, EP, Held, L, and Johnson, NA (2008). "Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds".

Gross, CW and Sohl, JE (1990). Disaggregation methods to expedite product line forecasting. *Journal of Forecasting* **9**(3), 233–254.

Hyndman, R (2017). forecast: Forecasting Functions for Time Series and Linear Models, R package version 8.0. *URL: http://github.com/robjhyndman/forecast*.

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

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

Jordan, A, Krüger, F, and Lerch, S (2017). Evaluating probabilistic forecasts with the R package scoringRules. arXiv: 1709.04743.

Kahn, KB (1998). *Revisiting top-down versus bottom-up forecasting*. http://search.ebscohost.com/login.aspx?direct=true%7B%5C&%7Ddb=bth%7B%5C&%7DAN=985713%7B%5C&%7Dlang=pt-br%7B%5C&%7Dsite=ehost-live.

Lapide, L (1998). A simple view of top-down vs bottom-up forecasting.pdf. *Journal of Business Forecasting Methods & Systems* **17**, 28–31.

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

Pinson, P and Tastu, J (2013). *Discrimination ability of the Energy score*. Tech. rep. Technical University of Denmark.

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

Schäfer, J and 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).

Scheuerer, M and Hamill, TM (2015). Variogram-Based Proper Scoring Rules for Probabilistic Forecasts of Multivariate Quantities *. *Monthly Weather Review* **143**(4), 1321–1334.

Schwarzkopf, AB, Tersine, RJ, and Morris, JS (1988). Top-down versus bottom-up forecasting strategies. *nternational Journal of Production Research* **26**(11), 1833.

Wickramasuriya, SL, Athanasopoulos, G, and Hyndman, RJ (2017). *Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization*. Working Paper 22/17. Department of Econometrics and Business Statistics, Monash University,Australia.

Yao, Q and Brockwell, PJ (2006). Gaussian maximum likelihood estimation for ARMA models. I. Time series. *Journal of Time Series Analysis* **27**(6), 857–875.