Probabilistic Forecasts in Hierarchical Time Series

January 9, 2018

1 Introduction.

- Introduction to hierarchical time series including a literature review in the context of point forecasts.
- Importance of probabilistic forecasts in time series.
- Lack of attention in the context of probabilistic forecasts in hierarchical literature.

2 Notations

We first introduce notations where possible we follow 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,

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

where $\boldsymbol{S}$ is a $n \times m$ constant matrix whose columns span the linear subspace for which all constraints hold. Since this linear subspace is equivalent to the column space of $\boldsymbol{S}$, we denote it as $\mathscr{C}(\boldsymbol{S})$. Further the null space of $\boldsymbol{S}$ which is orthogonal to this linear subspace is denoted by $\mathscr{N}(\boldsymbol{S})$. To understand the notations clearly, consider the hierarchy given in Figure 1.

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 $y_{T,t}$ denote the observation at level 0; $y_{A,t}, y_{B,t}$ denote observations at level 1; and $y_{AA,t}, y_{AB,t}, y_{AC,t}, y_{BA,t}, y_{BB,t}$ denote observations at level 2. Then $\boldsymbol{y}_t = [y_{T,t}, y_{A,t}, y_{B,t}, y_{C,t}, y_{AA,t}, y_{AB,t}, y_{AC,t}, y_{BA,t}, y_{BB,t}]^T$,

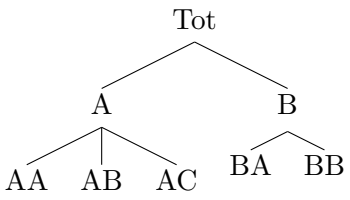

Figure 1: Two level hierarchical diagram

$\boldsymbol{b}_t = [y_{AA,t}, y_{AB,t}, y_{AC,t}, y_{BA,t}, y_{BB,t}]^T$, $m = 7$, n=11, and

$$\boldsymbol{S} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ & & I_5 & & \end{pmatrix},$$

where I_5 is a 5-dimension identity matrix.

3 Coherent forecasts

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

Definition 3.1: 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 it lies in a m-dimensional subspace of $\mathbb{R}^n$ which is spanned by the columns of $\boldsymbol{S}$.

Definition 3.2: Coherent Probabilistic Forecasts
Let $(\mathbb{R}^m, \mathscr{F}^m, \nu^m)$ be a probability measure space where $\mathscr{F}^m$ is a sigma algebra on $\mathbb{R}^m$. $\breve{\nu}(.)$ on $(\mathscr{C}(\boldsymbol{S}), \mathscr{F}_{\boldsymbol{S}})$ is said to be coherent probability measure 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 $\boldsymbol{S}$.

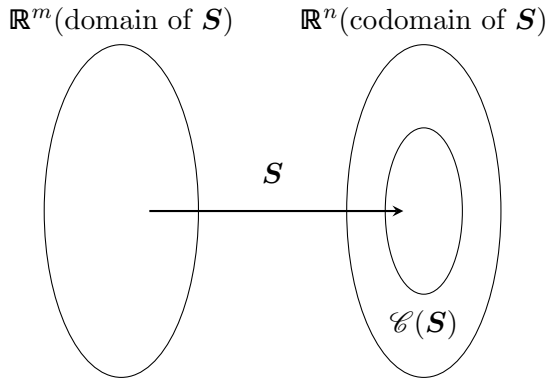

Figure 2: Any set $\boldsymbol{A} \in \mathbb{R}$ will be mapped to the $\mathscr{C}(\boldsymbol{S})$ through the mapping $\boldsymbol{S}$

Definition 3.2 implies the probability measure on $\mathscr{C}(\boldsymbol{S})$ is equivalent to the probability measure on $(\mathbb{R}^m, \mathscr{F}^m)$. Hence, there is no density anywhere outside the linear subspace spanned by the columns of $\boldsymbol{S}$. That is, a *Coherent probability density forecast* is any density $\boldsymbol{f}(\breve{\boldsymbol{y}}_{t+h})$ such that $\boldsymbol{f}(\breve{\boldsymbol{y}}_{t+h}) = 0$ for all $\breve{\boldsymbol{y}}_{t+h}$ in the null space of $\boldsymbol{S}$. 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 T. Suppose the forecasts at time $t+h$ of these series are given by $\breve{\boldsymbol{y}}_{t+h} = [\breve{y}_{T,t+h}, \breve{y}_{A,t+h}, \breve{y}_{B,t+h}]$ and $\breve{\boldsymbol{y}}_{t+h} \in \mathbb{R}^3$. Due to the aggregation constraint of the hierarchy we have $\breve{y}_{T,t+h} = \breve{y}_{A,t+h} + \breve{y}_{B,t+h}$. This implies, even though $\breve{\boldsymbol{y}}_{t+h}$ is in $\mathbb{R}^3$, the points actually lie in a two dimensional subspace within that $\mathbb{R}^3$ space. This subspace is equivalent to $\mathscr{C}(\boldsymbol{S})$ for this simple hierarchy. Therefore, for any $\breve{\boldsymbol{y}}_{t+h} \in \mathscr{N}(\boldsymbol{S})$ have a zero probability. I.e. $f(\breve{\boldsymbol{y}}_{t+h}) = 0$ for any $\breve{\boldsymbol{y}}_{t+h} \in \mathscr{N}(\boldsymbol{S})$.

By the definition of coherent forecasts, it is clear that there are only m number of linear independent series in the whole hierarchy. We refer to these as m dimensional *basis set of series* since these series generates a *basis* that spans the $\mathscr{C}(\boldsymbol{S})$. Other $n-m$ series of the hierarchy can be linearly determined by these basis set of series. This implies that any coherent density is a degenerate density. The m number of bottom level series of a given hierarchy can be considered as a basis set of series. Then the columns of $\boldsymbol{S}$ is a basis that generates through these bottom level series. This basis spans the linear subspace where the degenerate density lives on.

However bottom level series are not the only set of basis series and we can find many other basis set of series which generates basis that spans the same $\mathscr{C}(\boldsymbol{S})$ for a given hierarchy. If we go back to the hierarchy in example 1, instead of two bottom level series $(\boldsymbol{A}, \boldsymbol{B})$ we can take $(\boldsymbol{T}, \boldsymbol{A})$ as the basis set of series and the basis that generates through $(\boldsymbol{T}, \boldsymbol{A})$ will be $\{\begin{pmatrix}1 & 0 & 1\end{pmatrix}^T, \begin{pmatrix}0 & 1 & -1\end{pmatrix}^T\}$. Another possible basis would be the singular value decomposition of $\boldsymbol{S}$.

An important thing to notice here is all of these basis spans the same linear subspace equivalent to the $\mathscr{C}(\boldsymbol{S})$. Therefore the definition (3.2) is not unique and one can redefine the coherent probabilistic forecasts with respect to any basis. However we stick to the definition (3.2) and consider the basis defined by the columns of $\boldsymbol{S}$ that generates through the bottom levels of the hierarchy in what follows.

It is also worth to mention that the definition (3.1) and (3.2) 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. This is the only study that copes with probabilistic forecasts in hierarchical literature thus far.

4 Forecast reconciliation

Bottom-up, top-down and middle-out methods are the most traditional forecasting methods that were used to produce coherent point forecasts in early studies on hierarchical literature. In bottom-up approach, forecasts of the lowest level are first generated and they are simply aggregated to forecast the 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 proportions of observed data (Gross and Sohl, 1990). 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.

These three approaches use only a part of the information available when producing co-

herent forecasts. This might result inaccurate forecasts. For example, if the bottom level series are highly volatile or too noisy and hence challenging to forecast, then the resulting forecasts from bottom-up approach would be inaccurate.

As an alternative to these conventional methods, Hyndman et al. (2011) propose to use information from all the levels of the hierarchy and reconcile them to obtain coherent point forecasts. In this approach, independent forecasts of all series are initially obtained. It is very unlikely that these forecasts are coherent. Then, these forecasts are optimally combined through a regression model to obtain coherent forecasts. This study has given arise to the concept of point forecast reconciliation in hierarchical time series. In general, any forecasting method that use a set of incoherent forecasts and revise them to obtain coherent forecasts is referred to as forecast reconciliation method. For example, Minimum Trace reconciliation (Wickramasuriya, Athanasopoulos, and Hyndman, 2017), GTOP (Erven and Cugliari, 2014). It is important to notice that bottom-up, top-down and middle-out methods are not reconciliation methods since they only use forecasts from a part of the levels in the hierarchy when producing coherent forecasts.

4.1 Point forecast reconciliation

Even though the point forecast reconciliation has a well established literature, we would like to redefine this using concepts in linear algebra as a groundwork for 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

$$\tilde{\boldsymbol{b}}_{t+h} = \boldsymbol{g}(\hat{\boldsymbol{y}}_{t+h}), \tag{2}$$

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 $\tilde{\boldsymbol{y}}_{t+h} = \boldsymbol{S}(\tilde{\boldsymbol{b}}_{t+h})$ such that $\tilde{\boldsymbol{y}}_{t+h} \in \mathscr{C}(\boldsymbol{S})$, where $\boldsymbol{S}(\tilde{\boldsymbol{b}}_{t+h})$ is the image of $\tilde{\boldsymbol{b}}_{t+h}$ under $\boldsymbol{S}$ on the $\mathscr{C}(\boldsymbol{S}) < \mathbb{R}^n$.

In the following content we explain more on how this definition can be used in practice to reconcile point forecasts in hierarchical time series. Let $\boldsymbol{R} \in \mathbb{R}^{n \times n-m}$ consists the columns that spans $\mathscr{N}(\boldsymbol{S})$ which is orthogonal to $\mathscr{C}(\boldsymbol{S})$. $\mathscr{N}(\boldsymbol{S})$ is also equivalent to the $\mathscr{C}(\boldsymbol{R})$. Note that $\boldsymbol{R}$ is also not unique and one example is a matrix whose columns represent the aggregation constraints for the 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 follow the below 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 $\begin{pmatrix} \tilde{\boldsymbol{b}}_{t+h}^T & \vdots & \tilde{\boldsymbol{t}}_{t+h}^T \end{pmatrix}^T$ denote these coordinates. Note

that $\tilde{\boldsymbol{b}}_{t+h}$ is a basis set of series which is really the reconciled bottom level series that generates $\{\boldsymbol{s}_1, ..., \boldsymbol{s}_m\}$ and $\tilde{\boldsymbol{t}}_{t+h}$ is another basis set of series that generates $\{\boldsymbol{r}_1, ..., \boldsymbol{r}_{n-m}\}$. Then from basic properties of linear algebra it follows that,

$$\begin{pmatrix} \boldsymbol{S} & \vdots & \boldsymbol{R} \end{pmatrix} \begin{pmatrix} \tilde{\boldsymbol{b}}_{t+h}^T & \vdots & \tilde{\boldsymbol{t}}_{t+h}^T \end{pmatrix}^T = \hat{\boldsymbol{y}}_{t+h}, \tag{3}$$

$$\hat{\boldsymbol{y}}_{t+h} = \boldsymbol{S}\tilde{\boldsymbol{b}}_{t+h} + \boldsymbol{R}\tilde{\boldsymbol{t}}_{t+h}, \tag{4}$$

and

$$\begin{pmatrix} \tilde{\boldsymbol{b}}_{t+h}^T & \vdots & \tilde{\boldsymbol{t}}_{t+h}^T \end{pmatrix}^T = \begin{pmatrix} \boldsymbol{S} & \vdots & \boldsymbol{R} \end{pmatrix}^{-1} \hat{\boldsymbol{y}}_{t+h}. \tag{5}$$

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,

$$\begin{pmatrix} \boldsymbol{S} & \vdots & \boldsymbol{R} \end{pmatrix}^{-1} = \begin{pmatrix} (\boldsymbol{R}_\perp^T \boldsymbol{S})^{-1} \boldsymbol{R}_\perp^T \\ \cdots \\ (\boldsymbol{S}_\perp^T \boldsymbol{R})^{-1} \boldsymbol{S}_\perp^T \end{pmatrix}. \tag{6}$$

Thus we have,

$$\begin{pmatrix} \tilde{\boldsymbol{b}}_{t+h} \\ \cdots \\ \tilde{\boldsymbol{t}}_{t+h} \end{pmatrix} = \begin{pmatrix} (\boldsymbol{R}_\perp^T \boldsymbol{S})^{-1} \boldsymbol{R}_\perp^T \\ \cdots \\ (\boldsymbol{S}_\perp^T \boldsymbol{R})^{-1} \boldsymbol{S}_\perp^T \end{pmatrix} \hat{\boldsymbol{y}}_{t+h}. \tag{7}$$

From (7) it follows that,

$$\tilde{\boldsymbol{b}}_{t+h} = (\boldsymbol{R}_\perp^T \boldsymbol{S})^{-1} \boldsymbol{R}_\perp^T \hat{\boldsymbol{y}}_{t+h} \tag{8}$$

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 $\mathscr{C}(\boldsymbol{S})$ through $\boldsymbol{S}$. Thus we have,

$$\tilde{\boldsymbol{y}}_{t+h} = \boldsymbol{S}(\boldsymbol{R}_\perp^T \boldsymbol{S})^{-1} \boldsymbol{R}_\perp^T \hat{\boldsymbol{y}}_{t+h}, \quad \tilde{\boldsymbol{y}}_{t+h} \in \mathscr{C}(\boldsymbol{S}) < \mathbb{R}^n. \tag{9}$$

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 the mapping $\boldsymbol{g}$ in definition 4.1 is considered to be linear we have,

$$\boldsymbol{g} = (\boldsymbol{R}_\perp^T \boldsymbol{S})^{-1} \boldsymbol{R}_\perp^T. \tag{10}$$

In previous studies on hierarchical point forecasting, $\boldsymbol{g}$ is considered as a $m \times n$ matrix $\boldsymbol{P}$ and thus $\tilde{\boldsymbol{y}}_{t+h} = \boldsymbol{S}\boldsymbol{P}\hat{\boldsymbol{y}}_{t+h}$. In other words, $\boldsymbol{g}$ linearly projects incoherent point forecasts onto the $\mathscr{C}(\boldsymbol{S})$. Further in our context, we need to find $\boldsymbol{R}_\perp$ such that $\boldsymbol{R}_\perp^T \boldsymbol{S}$ is invertible. i.e, $(\boldsymbol{R}_\perp^T \boldsymbol{S})^{-1} \boldsymbol{R}_\perp^T \boldsymbol{S} = \boldsymbol{I}$. This condition coincides with the unbiased condition $\boldsymbol{S}\boldsymbol{P}\boldsymbol{S} = \boldsymbol{S}$ proposed by Hyndman et al. (2011).

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

$$\tilde{\boldsymbol{b}}_{t+h}^{OLS} = (\boldsymbol{S}^T\boldsymbol{S})^{-1}\boldsymbol{S}^T\hat{\boldsymbol{y}}_{t+h},$$

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

$$\tilde{\boldsymbol{y}}_{t+h}^{OLS} = \boldsymbol{S}(\boldsymbol{S}^T\boldsymbol{S})^{-1}\boldsymbol{S}^T\hat{\boldsymbol{y}}_{t+h}. \tag{11}$$

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 $\boldsymbol{R}_{\perp}^{T} = \boldsymbol{S}^T\hat{\boldsymbol{W}}_{T+h}^{-1}$ will minimize the trace of mean squared reconciled forecasts errors under the property of unbiasedness. $\hat{\boldsymbol{W}}_{T+h}^{-1}$ is the variance of the incoherent forecast errors. This will result,

$$\tilde{\boldsymbol{b}}_{t+h}^{MinT} = (\boldsymbol{S}^T\hat{\boldsymbol{W}}_{T+h}^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}^T\hat{\boldsymbol{W}}_{T+h}^{-1}\hat{\boldsymbol{y}}_{t+h},$$

and thus,

$$\tilde{\boldsymbol{y}}_{t+h}^{MinT} = \boldsymbol{S}(\boldsymbol{S}^T\hat{\boldsymbol{W}}_{T+h}^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}^T\hat{\boldsymbol{W}}_{T+h}^{-1}\hat{\boldsymbol{y}}_{t+h}. \tag{12}$$

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. $< \hat{\boldsymbol{y}}_{T+h}, \tilde{\boldsymbol{y}}_{T+h} >_{\hat{\boldsymbol{W}}}$.

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 measure space and $(\mathbb{R}^m, \mathscr{F}^m, \nu^m)$ be a probability measure space 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,

$$\nu^m(\boldsymbol{A}) = \hat{\nu}(\boldsymbol{g}^{-1}(\boldsymbol{A})), \quad \forall \quad \boldsymbol{A} \in \mathscr{F}^m. \tag{13}$$

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

$$\tilde{\nu}(\boldsymbol{S}(\boldsymbol{A})) = \hat{\nu}(\boldsymbol{g}^{-1}(\boldsymbol{A})), \quad \forall \quad \boldsymbol{A} \in \mathscr{F}^m, \tag{14}$$

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

Since the above definition seems not to be straight forward 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 (5). Suppose $\hat{\boldsymbol{f}}(.)$ is the probability density of $\hat{\boldsymbol{y}}_{t+h}$. Our goal is to reconcile $\hat{\boldsymbol{f}}(.)$ such that the density lives on the $\mathscr{C}(\boldsymbol{S})$. In order to obtain this reconciled density, we need to project $\hat{\boldsymbol{f}}(\hat{\boldsymbol{y}}_{t+h})$ onto the $\mathscr{C}(\boldsymbol{S})$ along the direction of

$\mathscr{C}(\boldsymbol{R})$.

Let the density of coordinates of $\hat{\boldsymbol{y}}_{t+h}$ with respect to basis $\boldsymbol{B}$ is denoted by $\boldsymbol{f_B}(.)$. Then it follows from (5) and the facts on density of transformed variables,

$$\boldsymbol{f_B}(\tilde{\boldsymbol{b}}_{t+h}, \tilde{\boldsymbol{t}}_{t+h}) = \hat{\boldsymbol{f}}(\boldsymbol{S}\tilde{\boldsymbol{b}}_{t+h} + \boldsymbol{R}\tilde{\boldsymbol{t}}_{t+h}) \quad \Big| \begin{array}{c|c|c} \boldsymbol{S} & \vdots & \boldsymbol{R} \end{array} \Big|, \tag{15}$$

where $|.|$ denote the determinant of a matrix. Now that we have the density of $\begin{pmatrix} \tilde{\boldsymbol{b}}_{t+h}^T & \vdots & \tilde{\boldsymbol{t}}_{t+h}^T \end{pmatrix}^T$, the marginal density of $(\tilde{\boldsymbol{b}}_{t+h})$ can be obtained by integrating (15) 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.,

$$\tilde{\boldsymbol{f}}(\tilde{\boldsymbol{b}}_{t+h}) = \int_{lim(\tilde{\boldsymbol{t}}_{t+h})} \hat{\boldsymbol{f}}(\boldsymbol{S}\tilde{\boldsymbol{b}}_{t+h} + \boldsymbol{R}\tilde{\boldsymbol{t}}_{t+h}) \quad \Big| \begin{array}{c|c|c} \boldsymbol{S} & \vdots & \boldsymbol{R} \end{array} \Big| \quad d\tilde{\boldsymbol{t}}_{t+h}. \tag{16}$$

Finally to get the reconciled density of the whole hierarchy, we simply follow the definition (3.2) and have,

$$\tilde{\boldsymbol{f}}(\tilde{\boldsymbol{y}}_{t+h}) = \boldsymbol{S} \circ \tilde{\boldsymbol{f}}(\tilde{\boldsymbol{b}}_{t+h}). \tag{17}$$

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

Example 2

Suppose $\hat{\boldsymbol{Y}}_{t+h} \sim \mathscr{N}(\hat{\boldsymbol{\mu}}_{t+h}, \hat{\boldsymbol{\Sigma}}_{t+h}) \leftrightarrow^d \hat{\boldsymbol{f}}(\hat{\boldsymbol{y}}_{t+h})$. Then from (15) it follows,

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

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

$$\begin{aligned} \boldsymbol{f_B}(.) &= \frac{\exp\left\{-\frac{1}{2}(\boldsymbol{S}\tilde{\boldsymbol{b}}_{t+h} + \boldsymbol{R}\tilde{\boldsymbol{t}}_{t+h} - \hat{\boldsymbol{\mu}}_{t+h})^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{t+h}}^{-1} (\boldsymbol{S}\tilde{\boldsymbol{b}}_{t+h} + \boldsymbol{R}\tilde{\boldsymbol{t}}_{t+h} - \hat{\boldsymbol{\mu}}_{t+h})\right\}}{(2\pi)^{\frac{n}{2}} \left|\hat{\boldsymbol{\Sigma}}_{t+h}\right|^{\frac{1}{2}} \left| \begin{array}{c|c|c} \boldsymbol{S} & \vdots & \boldsymbol{R} \end{array} \right|^{-1}}, \\ &= \frac{\exp\left\{-\frac{1}{2}\left(\left(\begin{array}{c|c|c} \boldsymbol{S} & \vdots & \boldsymbol{R} \end{array}\right)\begin{pmatrix} \tilde{\boldsymbol{b}}_{t+h}^T & \vdots & \tilde{\boldsymbol{t}}_{t+h}^T \end{pmatrix}^T - \hat{\boldsymbol{\mu}}_{t+h}\right)^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{t+h}}^{-1}\left(\left(\begin{array}{c|c|c} \boldsymbol{S} & \vdots & \boldsymbol{R} \end{array}\right)\begin{pmatrix} \tilde{\boldsymbol{b}}_{t+h}^T & \vdots & \tilde{\boldsymbol{t}}_{t+h}^T \end{pmatrix}^T - \hat{\boldsymbol{\mu}}_{t+h}\right)\right\}}{(2\pi)^{\frac{n}{2}} \left|\hat{\boldsymbol{\Sigma}}_{t+h}\right|^{\frac{1}{2}} \left| \begin{array}{c|c|c} \boldsymbol{S} & \vdots & \boldsymbol{R} \end{array} \right|^{-1}}, \end{aligned}$$

$$\begin{aligned} \boldsymbol{f_B}(.) = \frac{1}{(2\pi)^{\frac{n}{2}} \left|\hat{\boldsymbol{\Sigma}}_{t+h}\right|^{\frac{1}{2}} \left| \begin{array}{c|c|c} \boldsymbol{S} & \vdots & \boldsymbol{R} \end{array} \right|^{-1}} \exp\Big\{ -\frac{1}{2}\Big(&\begin{pmatrix} \tilde{\boldsymbol{b}}_{t+h}^T & \vdots & \tilde{\boldsymbol{t}}_{t+h}^T \end{pmatrix}^T - \begin{pmatrix} \boldsymbol{S} & \vdots & \boldsymbol{R} \end{pmatrix}^{-1} \hat{\boldsymbol{\mu}}_{t+h}\Big)^T \\ &\Big[\begin{pmatrix} \boldsymbol{S} & \vdots & \boldsymbol{R} \end{pmatrix} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{t+h}} \begin{pmatrix} \boldsymbol{S} & \vdots & \boldsymbol{R} \end{pmatrix}^T\Big]^{-1} \\ &\Big(\begin{pmatrix} \tilde{\boldsymbol{b}}_{t+h}^T & \vdots & \tilde{\boldsymbol{t}}_{t+h}^T \end{pmatrix}^T - \begin{pmatrix} \boldsymbol{S} & \vdots & \boldsymbol{R} \end{pmatrix}^{-1} \hat{\boldsymbol{\mu}}_{t+h}\Big)\Big\}. \end{aligned}$$

Recall that,

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

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

$$\boldsymbol{f_B}(.) = \frac{1}{(2\pi)^{\frac{n}{2}} \left| \hat{\boldsymbol{\Sigma}}_{t+h} \right|^{\frac{1}{2}} \left| \begin{pmatrix} \boldsymbol{P} \\ \boldsymbol{Q} \end{pmatrix} \right|} \exp\Big\{ -\frac{1}{2} \Big[\begin{pmatrix} \tilde{\boldsymbol{b}}_{t+h} \\ \tilde{\boldsymbol{t}}_{t+h} \end{pmatrix} - \begin{pmatrix} \boldsymbol{P} \\ \boldsymbol{Q} \end{pmatrix} \hat{\boldsymbol{\mu}}_{t+h} \Big]^T \Big[\begin{pmatrix} \boldsymbol{P} \\ \boldsymbol{Q} \end{pmatrix} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{t+h}} \begin{pmatrix} \boldsymbol{P} \\ \boldsymbol{Q} \end{pmatrix}^T \Big]^{-1} \Big[\begin{pmatrix} \tilde{\boldsymbol{b}}_{t+h} \\ \tilde{\boldsymbol{t}}_{t+h} \end{pmatrix} - \begin{pmatrix} \boldsymbol{P} \\ \boldsymbol{Q} \end{pmatrix} \hat{\boldsymbol{\mu}}_{t+h} \Big] \Big\},$$

$$\boldsymbol{f_B}(.) = \frac{1}{(2\pi)^{\frac{n}{2}} \left| \begin{pmatrix} \boldsymbol{P} \\ \boldsymbol{Q} \end{pmatrix} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{t+h}} \begin{pmatrix} \boldsymbol{P} \\ \boldsymbol{Q} \end{pmatrix}^T \right|^{\frac{1}{2}}} \exp\Big\{ -\frac{1}{2} \begin{pmatrix} \tilde{\boldsymbol{b}}_{t+h} - \boldsymbol{P}\hat{\boldsymbol{\mu}}_{t+h} \\ \tilde{\boldsymbol{t}}_{t+h} - \boldsymbol{Q}\hat{\boldsymbol{\mu}}_{t+h} \end{pmatrix}^T \Big[\begin{pmatrix} \boldsymbol{P} \\ \boldsymbol{Q} \end{pmatrix} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{t+h}} \begin{pmatrix} \boldsymbol{P} \\ \boldsymbol{Q} \end{pmatrix}^T \Big]^{-1} \begin{pmatrix} \tilde{\boldsymbol{b}}_{t+h} - \boldsymbol{P}\hat{\boldsymbol{\mu}}_{t+h} \\ \tilde{\boldsymbol{t}}_{t+h} - \boldsymbol{Q}\hat{\boldsymbol{\mu}}_{t+h} \end{pmatrix} \Big\}.$$

Since, $\Big[\begin{pmatrix} \boldsymbol{P} \\ \boldsymbol{Q} \end{pmatrix} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{t+h}} \begin{pmatrix} \boldsymbol{P} \\ \boldsymbol{Q} \end{pmatrix}^T \Big] = \begin{pmatrix} \boldsymbol{P}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t+h}}\boldsymbol{P}^T & \boldsymbol{P}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t+h}}\boldsymbol{Q}^T \\ \boldsymbol{Q}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t+h}}\boldsymbol{P}^T & \boldsymbol{Q}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t+h}}\boldsymbol{Q}^T \end{pmatrix}$ we have,

$$\boldsymbol{f_B}(.) = \frac{1}{(2\pi)^{\frac{n}{2}} \left| \begin{pmatrix} \boldsymbol{P}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t+h}}\boldsymbol{P}^T & \boldsymbol{P}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t+h}}\boldsymbol{Q}^T \\ \boldsymbol{Q}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t+h}}\boldsymbol{P}^T & \boldsymbol{Q}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t+h}}\boldsymbol{Q}^T \end{pmatrix} \right|^{\frac{1}{2}}} \exp\Big\{ -\frac{1}{2} \begin{pmatrix} \tilde{\boldsymbol{b}}_{t+h} - \boldsymbol{P}\hat{\boldsymbol{\mu}}_{t+h} \\ \tilde{\boldsymbol{t}}_{t+h} - \boldsymbol{Q}\hat{\boldsymbol{\mu}}_{t+h} \end{pmatrix}^T \Big[\begin{pmatrix} \boldsymbol{P}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t+h}}\boldsymbol{P}^T & \boldsymbol{P}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t+h}}\boldsymbol{Q}^T \\ \boldsymbol{Q}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t+h}}\boldsymbol{P}^T & \boldsymbol{Q}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t+h}}\boldsymbol{Q}^T \end{pmatrix} \Big]^{-1} \begin{pmatrix} \tilde{\boldsymbol{b}}_{t+h} - \boldsymbol{P}\hat{\boldsymbol{\mu}}_{t+h} \\ \tilde{\boldsymbol{t}}_{t+h} - \boldsymbol{Q}\hat{\boldsymbol{\mu}}_{t+h} \end{pmatrix} \Big\}.$$

$\boldsymbol{f_B}(.)$ gives the joint distribution of $\begin{pmatrix} \tilde{\boldsymbol{b}}_{t+h}^T & \vdots & \tilde{\boldsymbol{t}}_{t+h}^T \end{pmatrix}^T$, which is a multivariate Gaussian distribution. Then from (16) and the properties of marginalization of multivariate Gaussian distribution it follows,

$$\tilde{\boldsymbol{f}}(\tilde{\boldsymbol{b}}_{t+h}) = \frac{1}{(2\pi)^{\frac{n}{2}} \left| \boldsymbol{P}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t+h}}\boldsymbol{P}^T \right|^{\frac{1}{2}}} \exp\Big\{ -\frac{1}{2} (\tilde{\boldsymbol{b}}_{t+h} - \boldsymbol{P}\hat{\boldsymbol{\mu}}_{t+h})^T (\boldsymbol{P}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t+h}}\boldsymbol{P}^T)^{-1} (\tilde{\boldsymbol{b}}_{t+h} - \boldsymbol{P}\hat{\boldsymbol{\mu}}_{t+h}) \Big\}. \tag{18}$$

(18) implies $\tilde{\boldsymbol{b}}_{t+h} \sim \mathcal{N}(\boldsymbol{P}\hat{\boldsymbol{\mu}}_{t+h}, \boldsymbol{P}\hat{\boldsymbol{\Sigma}}_{t+h}\boldsymbol{P}^T)$ where $\boldsymbol{P} = (\boldsymbol{R}_\perp^T \boldsymbol{S})^{-1} \boldsymbol{R}_\perp^T$. Then from (17) it follows that,

$$\tilde{\boldsymbol{f}}(\tilde{\boldsymbol{y}}_{t+h}) = \tilde{\boldsymbol{f}}(\boldsymbol{S}\tilde{\boldsymbol{b}}_{t+h}). \tag{19}$$

Therefore, the reconciled Gaussian forecast distribution of the whole hierarchy is $\mathcal{N}(\boldsymbol{SP}\hat{\boldsymbol{\mu}}_{t+h}, \boldsymbol{SP}\hat{\boldsymbol{\Sigma}}_{t+h}\boldsymbol{P}^T\boldsymbol{S}^T$

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,

$$E_{\boldsymbol{G}}[S(\boldsymbol{Y}, \boldsymbol{y})] \leq E_{\boldsymbol{G}}[S(\breve{\boldsymbol{Y}}, \boldsymbol{y})], \tag{20}$$

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

Following table summarizes few existing proper scoring rules.

Table 1: *Scoring rules to evaluate multivariate forecast densities*

Scoring rule	**Expression**	**Reference**
Log score	$LS(\breve{\boldsymbol{F}}, \boldsymbol{y}_{T+h}) = -log\breve{\boldsymbol{f}}(\boldsymbol{y}_{T+h})$	Gneiting and Raftery (2007)
Energy score	$eS(\breve{\boldsymbol{Y}}_{\boldsymbol{T+h}}, \boldsymbol{y}_{\boldsymbol{T+h}}) = E_{\breve{\boldsymbol{F}}}\|\|\breve{\boldsymbol{Y}}_{T+h} - \boldsymbol{y}_{T+h}\|\|^{\alpha} - \frac{1}{2}E_{\breve{\boldsymbol{F}}}\|\|\breve{\boldsymbol{Y}}_{T+h} - \breve{\boldsymbol{Y}}'_{T+h}\|\|^{\alpha}, \alpha \in (0, 2]$	Gneiting et al. (2008)
Variogram score	$VS(\breve{\boldsymbol{F}}, \boldsymbol{y}_{T+h}) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(\|y_{T+h,i} - y_{T+h,j}\|^p - E_{\breve{\boldsymbol{F}}}\|\breve{Y}_{T+h,i} - \breve{Y}_{T+h,j}\|^p\right)^2$	Scheuerer and Hamill (2015)

NOTE: $\breve{\boldsymbol{Y}}_{T+h}$ *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* i^{th} *and* j^{th} *components of the vector* $\breve{\boldsymbol{Y}}_{T+h}$*. Further the variogram score given is for order* p *where,* w_{ij} *are non-negative weights.*

Even though the log score can be used evaluate simulated forecast densities with large samples (Jordan, Krger, 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 about this in the next sub section.

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

$$eS(\breve{\boldsymbol{Y}}_{\boldsymbol{T+h}}, \boldsymbol{y}_{\boldsymbol{T+h}}) = ||\boldsymbol{y}_{T+h} - \breve{\boldsymbol{\mu}}_{T+h}||^2, \tag{21}$$

where $\breve{\boldsymbol{\mu}}_{T+h} = E_{\boldsymbol{F}}(\breve{\boldsymbol{Y}}_{T+h})$. 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 $\breve{\boldsymbol{F}}$, the variogram score is defined as,

$$VS(\breve{\boldsymbol{F}}, \boldsymbol{y}_{T+h}) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(|y_{T+h,i} - y_{T+h,j}|^p - \frac{1}{m}\sum_{k=1}^{B}|\breve{Y}^k_{T+h,i} - \breve{Y}^k_{T+h,j}|^p\right)^2. \tag{22}$$

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:

$$E_f[S(f, f)] = E_f[S(p_1, f)] = -ln[p_1(x = 0)] = 0,$$

and that of p_2 is:

$$E_f[S(p_2, f)] = -ln[p_2(x = 0)] < 0.$$

Therefore $S(f, f) > S(p_2, f)$ and hence there exist at least one forecast density which breaks the condition (20) for proper scoring rule. This implies log score cannot be used to evaluate the degenerate densities.

This suggest that 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) for evaluating univariate forecast densities would be helpful for this.

$$CRPS(\breve{F}_i, y_{T+h,i}) = E_{\breve{F}_i}|\breve{Y}_{T+h,i} - y_{T+h,i}| - \frac{1}{2}E_{\breve{F}_i}|\breve{Y}_{T+h,i} - \breve{Y}'_{T+h,i}|, \tag{23}$$

where $\breve{Y}_{T+h,i}$ and $\breve{Y}'_{T+h,i}$ are two independent copies from the i^{th} reconciled marginal forecast distribution $\tilde{F}_i$ of the hierarchy and $y_{T+h,i}$ is the i^{th} realization from the true marginal distribution G_i.

5.3 Simulation study

6 Conclusions.

References

Ben Taieb, S., Huser, R., Hyndman, R. J., and Genton, M. G. (2017). "Forecasting uncertainty in electricity smart meter data by boosting additive quantile regression". In: *IEEE Transactions on Smart Grid* 7.5, pp. 2448–2455.

Dunn, D. M., Williams, W. H., and Dechaine, T. L. (1976). "Aggregate Versus Subaggregate Models in Local Area Forecasting". In: *Journal of American Statistical Association* 71.353, pp. 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.

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

Gneiting, T. and Raftery, A. E. (2007). "Strictly Proper Scoring Rules, Prediction, and Estimation". In: *Journal of the American Statistical Association* 102.477, pp. 359–378.

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L., and Johnson, N. A. (2008). "Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds".

Gross, C. W. and Sohl, J. E. (1990). "Disaggregation methods to expedite product line forecasting". In: *Journal of Forecasting* 9.3, pp. 233–254.

Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., and Shang, H. L. (2011). "Optimal combination forecasts for hierarchical time series". In: *Computational Statistics and Data Analysis* 55.9, pp. 2579–2589.

Jordan, A., Krger, F., and Lerch, S. (2017). "Evaluating probabilistic forecasts with the R package scoringRules." In: URL: http://arxiv.org/abs/1709.04743http://arxiv.org/abs/1709.04743.

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

Scheuerer, M. and Hamill, T. M. (2015). "Variogram-Based Proper Scoring Rules for Probabilistic Forecasts of Multivariate Quantities *". In: *Monthly Weather Review* 143.4, pp. 1321–1334.

Wickramasuriya, S. L., Athanasopoulos, G., and Hyndman, R. J. (2017). "Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization".