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

21 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 (2018) 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.

Previous studies on coherent point forecasting have shown that reconciliation provides better coherent forecasts than the traditional bottom-up and top-down methods (Hyndman et al., 2011; Erven and Cugliari, 2014; Wickramasuriya, Athanasopoulos, and Hyndman, 2018). However, this idea has not been explored in the context of probabilistic forecasting.

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 forecast 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 largely follows that introduced in Wickramasuriya, Athanasopoulos, and Hyndman (2018). Suppose $\boldsymbol{y}_t \in \mathbb{R}^n$ comprises all observations of the hierarchy at time t and $\boldsymbol{b}_t \in \mathbb{R}^m$ comprises only the bottom-level observations at time t. Then due to the aggregation nature of the hierarchy we have

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

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

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

Consider the hierarchy given in Figure 1.

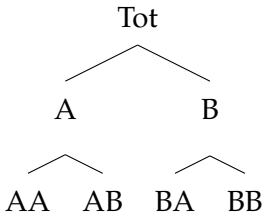

Figure 1: *Two level hierarchical diagram.*

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

While coherent point forecasts have been discussed many times previously, the definitions of coherence previously given are vague and are not easily extended to the situation of probabilistic forecasting.

We first give a new definition for coherent point forecasts using the properties of vector spaces, and then provide a definition of coherent probabilistic forecasts.

Definition 3.1 (Coherent subspace) *Suppose an n-dimensional time series $\boldsymbol{y}_t \in \mathbb{R}^n$ is subject to the linear aggregation constraint $\boldsymbol{y}_t = \boldsymbol{S}\boldsymbol{b}_t$, where $\boldsymbol{b}_t \in \mathbb{R}^m$ and $\boldsymbol{S}$ is an $n \times m$ constant matrix. Let $\mathbb{C}^m$ be an m-dimensional subspace of $\mathbb{R}^n$, where $\mathbb{C}^m < \mathbb{R}^n$. Then $\mathbb{C}^m$ is said to be a coherent space if it is spanned by the columns of $\boldsymbol{S}$.*

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 equivalent to the null space of $\boldsymbol{S}$, which we denote by $\mathbb{N}^{n-m}$.

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

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

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

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

We illustrate Definition 3.3 in Figure 2, showing $\boldsymbol{S}$ mapping any set $\boldsymbol{A} \in \mathbb{R}^m$ to the coherent space.

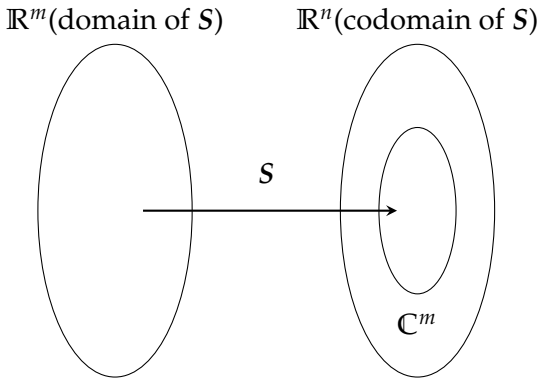

Figure 2: *Any set* $\boldsymbol{A} \in \mathbb{R}^m$ *will be mapped to* $\mathbb{C}^m$ *through the mapping* $\boldsymbol{S}$*.*

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 $\boldsymbol{f}(\breve{\boldsymbol{y}}_{t+h})$ such that $\boldsymbol{f}(\breve{\boldsymbol{y}}_{t+h}) = 0$ for all $\breve{\boldsymbol{y}}_{t+h} \in \mathbb{N}^{n-m}$.

The following example will help to understand these definitions more clearly.

Consider a simple hierarchy with two bottom-level series A and B that add up to the top level series Tot. Suppose the forecasts of these series at time $t+h$ are given by $\breve{\boldsymbol{y}}_{t+h} = [\breve{y}_{Tot,t+h}, \breve{y}_{A,t+h}, \breve{y}_{B,t+h}]$. 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 that, even though $\breve{\boldsymbol{y}}_{t+h} \in \mathbb{R}^3$, the points actually lie in $\mathbb{C}^2$, which is a two dimensional subspace within that $\mathbb{R}^3$ space. Therefore, any $\breve{\boldsymbol{y}}_{t+h} \in \mathbb{N}$ is impossible, so that $f(\breve{\boldsymbol{y}}_{t+h}) = 0$ 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 small hierarchy considered above, $\{(1\ 1\ 0)', (1\ 0\ 1)'\}$, $\{(1\ 0\ 1)', (0\ 1\ \ -1)'\}$ and the singular value decomposition of these two are some alternative basis vectors that span 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 those basis vectors. We refer to the coefficients of these linear combinations as the *basis series*. It is apparent that these basis series are m-dimensional and linearly independent in a given hierarchy. For example, in the smallest hierarchy, (A, B) and (Tot, A) are the basis series corresponding to the basis vectors $\{(1\ 1\ 0)', (1\ 0\ 1)'\}$, $\{(1\ 0\ 1)', (0\ 1\ \ -1)'\}$ respectively. Thus it is clear that the set of bottom-level series is a basis series that corresponds to the column vectors of $\boldsymbol{S}$.

Because the basis is not unique for a given coherent subspace, Definition 3.3 is not unique, and one can redefine the coherent probabilistic forecasts with respect to any basis. However, we stick to Definition 3.3 and consider the basis defined by the columns of $\boldsymbol{S}$ in what follows.

Definitions 3.2 and 3.3 facilitate extension to probabilistic forecast reconciliation which we discuss in the next section. In contrast to our definition, Ben Taieb et al. (2017) define coherent probabilistic forecasts in terms of convolutions. According to their definition, if the forecasts are coherent, then the convolution of forecast distributions of disaggregate series is identical to the forecast distribution of the corresponding aggregate series. While this is consistent with our definition, it is not easy to extend their definition to deal with probabilistic reconciliation.

4 Forecast reconciliation

Initially we define point forecast reconciliation, before extending the idea to the probabilistic setting.

4.1 Point forecast reconciliation

Definition 4.1 *Let $\hat{\boldsymbol{y}}_{t+h} \in \mathbb{R}^n$ be any set of incoherent forecasts at time $t+h$, and let*

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

where $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^m$ and $\boldsymbol{S} \circ \boldsymbol{g}(\cdot)$ is a projection of $\boldsymbol{g}(\cdot)$ onto $\mathbb{C}^m$. Then $\tilde{\boldsymbol{y}}_{t+h}$ is said to be "reconciled" if $\tilde{\boldsymbol{y}}_{t+h} \in \mathbb{C}^m$.

Definition 4.1 allows for both linear and non-linear reconciliation. In other words, if g is a non-linear function, then the reconciliation of $\hat{\boldsymbol{y}}_{t+h}$ will be non-linear, while if g is a linear function, then $\boldsymbol{S} \circ g(\cdot)$ will linearly project incoherent point forecasts onto $\mathbb{C}^m$. Previous studies in hierarchical point forecasting have only focussed on the linear case, $g(\boldsymbol{y}) = \boldsymbol{P}\boldsymbol{y}$, where $\boldsymbol{P}$ is an $m \times n$ matrix, and so $\tilde{\boldsymbol{y}}_{t+h} = \boldsymbol{S}\boldsymbol{P}\hat{\boldsymbol{y}}_{t+h}$.

Using Definition 4.1, we can now explain previous results for linear forecast reconciliation. Let $\boldsymbol{R} \in \mathbb{R}^{n\times(n-m)}$ comprise the columns that span $\mathbb{N}^{n-m}$, which is orthogonal to $\mathbb{C}^m$. Note that $\boldsymbol{R}$ is not unique; one example is a matrix whose columns represent the aggregation constraints for a given hierarchy. 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 $\{\boldsymbol{s}_1, \dots, \boldsymbol{s}_m\}$ and $\{\boldsymbol{r}_1, \dots, \boldsymbol{r}_{n-m}\}$ denote the columns of $\boldsymbol{S}$ and $\boldsymbol{R}$ respectively. Then $\boldsymbol{B} = \{\boldsymbol{s}_1, \dots, \boldsymbol{s}_m, \boldsymbol{r}_1, \dots, \boldsymbol{r}_{n-m}\}$ 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 $(\tilde{\boldsymbol{b}}'_{t+h}, \tilde{\boldsymbol{t}}'_{t+h})'$ denote these coordinates. Note that $\tilde{\boldsymbol{b}}_{t+h}$ is a basis series which is equivalent to the reconciled bottom-level series, and which corresponds to the linear combination coefficients of the basis $\{\boldsymbol{s}_1, \dots, \boldsymbol{s}_m\}$. Similarly, $\tilde{\boldsymbol{t}}_{t+h}$ is another basis series corresponding to the linear combination coefficients of the basis $\{\boldsymbol{r}_1, \dots, \boldsymbol{r}_{n-m}\}$. Then from basic properties of linear algebra it follows that

$$(\boldsymbol{S} \vdots \boldsymbol{R})(\tilde{\boldsymbol{b}}'_{t+h},\ \tilde{\boldsymbol{t}}'_{t+h})' = \hat{\boldsymbol{y}}_{t+h},$$

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

and

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

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

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

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 from the definition for coherent forecasts. 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,

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

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

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).

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 are given by

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

They referred this to as the 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 (2018), selecting $\boldsymbol{R}_{\perp}' = \boldsymbol{S}'\boldsymbol{W}_h^{-1}$ 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 in

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

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 the MinT solution. 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

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

Definition 4.2 *Suppose* $(\mathbb{R}^n, \mathscr{F}^n, \hat{\nu})$ *is an incoherent probability triple and* $(\mathbb{R}^m, \mathscr{F}^m, \nu^m)$ *is a probability triple defined on* $\mathbb{R}^m$*. Let* $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^m$*. Then the probability measure on the 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 probability measure of the whole reconciled hierarchy is given by

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

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

We now discuss how this definition 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 $\hat{\boldsymbol{y}}_{t+h}$ with respect to the basis $\boldsymbol{B}$ are given by (1). Suppose $\hat{\boldsymbol{f}}(\cdot)$ is the probability density of $\hat{\boldsymbol{y}}_{t+h}$. Our goal is to reconcile $\hat{\boldsymbol{f}}(\cdot)$ such that the density lives on $\mathbb{C}^m$. In order to obtain this reconciled density, we need to project $\hat{\boldsymbol{f}}(\hat{\boldsymbol{y}}_{t+h})$ onto $\mathbb{C}^m$ along the direction of $\mathbb{N}^{n-m}$.

Let the density of $\hat{\boldsymbol{y}}_{t+h}$ with respect to basis $\boldsymbol{B}$ be denoted by $f_{\boldsymbol{B}}(\cdot)$. Then it follows from (1), and standard results for densities of transformed variables, that

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

where $|\cdot|$ denotes the determinant of a matrix. Now that we have the density of $(\tilde{\boldsymbol{b}}'_{t+h}, \tilde{\boldsymbol{t}}'_{t+h})'$, 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 in the reconciled density of the bottom-level series $\tilde{\boldsymbol{b}}_{t+h}$,

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

Finally to get the reconciled density of the whole hierarchy, we simply follow Definition 3.3 to obtain

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

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

Example 2

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

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

By substituting the Gaussian distribution function for $f_{\boldsymbol{B}}(\cdot)$ we get

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

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

Recall that

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

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

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

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

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

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

This is the joint multivariate Gaussian distribution of $(\tilde{\boldsymbol{b}}'_{t+h} \vdots \tilde{\boldsymbol{t}}'_{t+h})'$. Then from (4) and the properties of the multivariate Gaussian distribution, it follows that

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

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

$$\tilde{f}(\tilde{\boldsymbol{y}}_{t+h}) = \tilde{f}(\boldsymbol{S}\tilde{\boldsymbol{b}}_{t+h}). \quad (7)$$

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

5 Evaluation of hierarchical probabilistic forecasts

The necessary final step in hierarchical forecasting is to make sure that our forecast distributions are accurately predicting the uncertain future. In general, forecasters prefer to maximize the sharpness of the forecast 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 forecast distribution should be equivalent in distribution to the realizations. On the other hand, sharpness refers to the spread or the concentration of the prediction distributions and it is a property of the forecasts only. The more concentrated the forecast distributions, the sharper the forecasts (Gneiting et al., 2008). However, independently assessing the calibration and sharpness will not help to properly evaluate the probabilistic forecasts. Therefore we need to assess these properties simultaneously using scoring rules.

Scoring rules are summary measures obtained based on the relationship between the forecast distributions and the realizations. In some studies, researchers take the scoring rules to be positively oriented, in which case the scores should be maximized (Gneiting and Raftery, 2007). However, scoring rules have also been defined to be negatively oriented, and then the scores should be minimized (Gneiting and Katzfuss, 2014). We consider negatively oriented scoring rules to evaluate probabilistic forecasts in hierarchical time series.

Let $\check{\boldsymbol{Y}}$ and $\boldsymbol{Y}$ be n-dimensional random vectors from the forecast distribution $\boldsymbol{F}$ and the true distribution $\boldsymbol{G}$, respectively. Further let $\boldsymbol{y}$ be an n-dimensional realization from $\boldsymbol{G}$. Then a scoring rule is a numerical value $S(\check{\boldsymbol{Y}}, \boldsymbol{y})$ assigned to each pair $(\check{\boldsymbol{Y}}, \boldsymbol{y})$. It is a "proper" scoring rule if

$$\mathrm{E}_{\boldsymbol{G}}[S(\boldsymbol{Y}, \boldsymbol{y})] \leq \mathrm{E}_{\boldsymbol{G}}[S(\check{\boldsymbol{Y}}, \boldsymbol{y})], \tag{8}$$

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 a few existing proper scoring rules.

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

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

Even though the log score can be used to evaluate simulated forecast densities with large samples (Jordan, Krüger, and Lerch, 2017), it is more convenient to use if we can assume a parametric forecast density for the hierarchy. However, the degeneracy of coherent forecast densities is problematic when using log scores. We will discuss this further in the next subsection.

For the energy score with $\alpha = 2$, it can be easily shown that

$$\mathrm{ES}(\boldsymbol{Y}_{T+h}, \check{y}_{T+h}) = \lVert\boldsymbol{y}_{T+h} - \check{\boldsymbol{\mu}}_{T+h}\rVert^2, \tag{9}$$

where $\check{\boldsymbol{\mu}}_{T+h} = \mathrm{E}_{\boldsymbol{F}}(\check{\boldsymbol{Y}}_{T+h})$. Therefore, in the limiting case, the energy score only measures the accuracy of the forecast mean, and not the entire distribution. Similarly, 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.

In contrast, Scheuerer and Hamill (2015) have shown that the variogram score has a higher discrimination ability of misspecified means, variances and correlation structures than the energy score. For a finite sample of size B from the multivariate forecast density $\breve{\boldsymbol{F}}$, the empirical variogram score is defined as

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

Scheuerer and Hamill (2015) recommend using $p = 0.5$.

5.1 Evaluating coherent forecast densities

Any coherent hierarchical forecast density is a degenerate density. To the best of our knowledge, there is no proper multivariate scoring rule available to evaluate degenerate densities. Further, it can easily be seen that some of the existing scoring rules breakdown under degeneracy.

For example, consider 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 $p_1(x)$ be equivalent to the true density, i.e. $p_1(x) = \mathbb{1}\{x = 0\}$, and let $p_2(x) \sim \mathcal{N}(0, \sigma^2)$ with $\sigma^2 < (2\pi)^{-1}$. The expected log score of p_1 is

I think you need Dirac delta functions here.

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

while 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 there exists at least one forecast density which breaks the condition (8) for a proper scoring rule. This implies that the log score cannot be used to evaluate the degenerate densities.

However, 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 the basis set of series, then we have almost obtained the correct forecast distribution of the whole hierarchy. Therefore, we propose to evaluate the forecast distributions using only the basis set of series. Then we can use any of the multivariate scoring rules discussed above without incurring problems due to degeneracy.

For example, since the bottom-level series forms a set of basis series for a given hierarchy, we can evaluate the coherent forecast distribution using only the bottom-level series instead of evaluating the whole distribution. In particular, if our purpose is to compare two coherent forecast densities, we can compare them using 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 metric spaces.

However we could compare the individual margins of the forecast density of the hierarchy using univariate proper scoring rules. The widely used Continuous Ranked Probability Score (CRPS) could then be used. This is defined as

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

where $\breve{Y}_{T+h,i}$ and $\breve{Y}^*_{T+h,i}$ 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 could also use univariate log scores, for which we could assume a parametric forecast distribution.

6 Probabilistic forecast reconciliation in the Gaussian framework

An important special case for probabilistic forecasting arises when we can assume a multivariate Gaussian distribution. That is, suppose all the historical data in the hierarchy follows a multivariate Gaussian distribution, $\boldsymbol{y}_T \sim \mathcal{N}(\boldsymbol{\mu}_T, \boldsymbol{\Sigma}_T)$, where both $\boldsymbol{\mu}_T$ and $\boldsymbol{\Sigma}_T$ live in $\mathbb{C}^m$ by nature of the hierarchical structure of the data. 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 live in $\mathbb{C}^m$.

It is well known that the optimal point forecasts with respect to the minimal mean square error are given by the conditional expectations, $\text{E}[Y_{T+h,i} \mid y_{1,i}, \ldots, y_{T,i}]$, $i = 1, \ldots, n$. Suppose we independently fit time series models for each series in the hierarchy. Then the point forecasts, $\hat{Y}_{T+h,i}$,

from the estimated models are unbiased and consistent estimators of $\mathrm{E}[Y_{T+h,i} \mid y_{1,i},\dots,y_{T,i}]$, assuming the parameter estimates of the fitted models are unbiased and asymptotically consistent.

Only for linear models?

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

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

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

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

Since $\boldsymbol{\alpha} = (\alpha_1,\dots,\alpha_p)'$ and $\boldsymbol{\beta} = (\beta_1,\dots,\beta_q)'$ 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 $\mathrm{E}[Y_{T+h,i} \mid y_{1,i},\dots,y_{T,i}]$. i.e.,

$$\hat{Y}_{T+h,i} \xrightarrow{p} \mathrm{E}[Y_{T+h,i} \mid y_{1,i},\dots,y_{T,i}] \quad \text{as} \quad T \to \infty. \tag{10}$$

Let $\hat{\boldsymbol{Y}}_{T+h} = (\hat{Y}_{T+h,1},\dots,\hat{Y}_{T+h,n})'$ and suppose (10) holds for $i = 1,\dots,n$. Then from Slutsky's theorem it follows that

$$\hat{\boldsymbol{Y}}_{T+h} \xrightarrow{p} \mathrm{E}[\boldsymbol{Y}_{T+h}|\boldsymbol{\mathcal{I}}_T] \quad \text{as} \quad T \to \infty. \tag{11}$$

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

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

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

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

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

From (11) it immediately follows that

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

That is,

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

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

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 do not 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

$$\widehat{\boldsymbol{Y}_{T+h,i}|\boldsymbol{\mathcal{I}}_T} \sim \mathcal{N}(\hat{\boldsymbol{Y}}_{T+h}, \boldsymbol{W}_{T+h}) \tag{12}$$

Since our primary objective is to find the coherent forecast density of the hierarchy, we need to reconcile (12). Using (7), the reconciled Gaussian forecast distribution is then given by

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

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

Result 1: Choosing $\boldsymbol{R}_\perp' = \boldsymbol{S}'\boldsymbol{W}_{T+h}^{-1}$ will ensure that 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 (9), the energy score at the upper limit of $\alpha = 2$ is given by $\|\boldsymbol{y}_{T+h} - \boldsymbol{S}\boldsymbol{P}\hat{\boldsymbol{y}}_{T+h}\|^2$. Then the expectation of the energy score with respect to the true distribution is equivalent to the trace of mean squared forecast error; i.e.,

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

From Theorem 1 of Wickramasuriya, Athanasopoulos, and Hyndman (2018) it immediately follows that $\boldsymbol{P} = (\boldsymbol{S}'\boldsymbol{W}_{T+h}^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}_{T+h}^{-1}$ minimizes the expected energy score, if we constrain the reconciled forecasts to be unbiased. Thus we have $\boldsymbol{R}_\perp' = \boldsymbol{S}'\boldsymbol{W}_{T+h}^{-1}$.

Table 2: *Several possible 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 more 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 (2018), where Diag($\boldsymbol{A}$) denotes the diagonal matrix of $\boldsymbol{A}$, $\tau = \frac{\sum_{i\neq j}\hat{Var}(\hat{r}_{ij})}{\sum_{i\neq j}\hat{r}^2_{ij}}$, and $\hat{r}_{ij}$ is the ijth element of the sample correlation matrix.*

Method	**Estimate of $\boldsymbol{W}_h$**	**Estimate of $\boldsymbol{R}'_{\perp}$**
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)	$\hat{\boldsymbol{W}}^{wls}_{T+1} = \text{Diag}(\hat{\boldsymbol{W}}^{shr}_{T+1})$	$\boldsymbol{S}'(\hat{\boldsymbol{W}}^{wls}_{T+1})^{-1}$

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

All of these forecasting methods are well-established in the context of point forecast reconciliation (Hyndman et al., 2011; Hyndman, Lee, and Wang, 2016; Wickramasuriya, Athanasopoulos, and Hyndman, 2018). Here, we are showing how these reconciliation methods can be used in the context of probabilistic forecast reconciliation, at least in the Gaussian framework.

Simulations

We consider the hierarchy given in Figure 1, comprising two aggregation levels with four bottom-level series. Each bottom-level series will be generated first, and then summed to obtain the data for the upper-level series. In practice, hierarchical time series tend to contain much noisier series at lower levels of aggregation. In order to replicate this feature in our simulations, we follow the data generating process proposed by Wickramasuriya, Athanasopoulos, and Hyndman (2018).

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 $\{\varepsilon_{AA,t}, \varepsilon_{AB,t}, \varepsilon_{BA,t}, \varepsilon_{BB,t}\} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$. The parameters for the 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:

$$y_{AA,t} = w_{AA,t} + u_t - 0.5v_t,$$
$$y_{AB,t} = w_{AB,t} - u_t - 0.5v_t,$$

$$y_{BA,t} = w_{BA,t} + u_t + 0.5v_t,$$
$$y_{BB,t} = w_{BB,t} - u_t + 0.5v_t,$$

where $u_t \sim \mathcal{N}(0, \sigma_u^2)$ and $v_t \sim \mathcal{N}(0, \sigma_v^2)$. To obtain the aggregate series at level 1, we add the bottom-level series:

$$y_{A,t} = w_{AA,t} + w_{AB,t} - v_t,$$
$$y_{B,t} = w_{BA,t} + w_{BB,t} + v_t,$$

and the total series is obtained using

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

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

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

Therefore

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

where $\boldsymbol{l}_1 = (1,1,1,1)'$, $\boldsymbol{l}_2 = (1,1,0,0)'$ and $\boldsymbol{l}_3 = (1,0,0,0)'$, and hence

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

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

We generate data for the hierarchy with sample size $T = 501$. Univariate ARIMA models were fitted for each series independently using the first 500 observations, and 1-step ahead base (incoherent) forecasts were calculated. We use the *forecast* package (Hyndman, 2017) in R (R Core Team, 2018) for model fitting and forecasting. The different estimates of $\boldsymbol{W}_{T+1}$ and the corresponding $\boldsymbol{R}_\perp'$ from Table 2 were obtained. This process was replicated using 1000 different data sets from the same data generating processes.

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. Entries in these columns show the percentage increase of score for different reconciliation methods 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	

To assess the predictive performance of different forecasting methods, we use scoring rules as discussed in Section 5. To faciliate comparisons, we report skill scores (Gneiting and Raftery, 2007). For a given forecasting method, evaluated by a particular scoring rule $S(\cdot)$, the skill score is calculated as

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

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

In Table 3, we compare different reconciliation methods over the conventional bottom-up method. 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.

We also evaluate the predictive ability of coherent forecasts over incoherent forecasts in Tables 4 and 5. Here 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.

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) out-perform the bottom-up forecasts. Further, these two methods produce probabilistic forecasts with the 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 4: *Comparison of incoherent vs coherent forecasts for the aggregate series using Skill scores. The "Incoherent" row shows the average scores for incoherent forecasts. Each entry above this row represents the percentage skill score with reference to the incoherent forecasts. Entries show the percentage increase in score for different forecasting methods relative to the 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*

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

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, 2018) 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.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

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 (2018). Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *J American Statistical Association*. to appear.

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.