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

April 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

27 April 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 Hierarchical Time Series

In the section, and throughout the paper, we will try to follow notational conventions used in Wickramasuriya, Athanasopoulos, and Hyndman (2018) as much as possible. A *hierarchical time series* is a collection of n variables where some variables are aggregates of other variables. For example in the hierarchy depicted in Figure 1 below, the variable labelled *Tot* is the sum of the series A and series B, the series A is the sum of series AA and series AB and the series B is the sum of the series BA and BB. The *bottom level series* are defined as those m variables that cannot be formed as aggregates of other variables, in the example in Figure 1 these are the series AA, AB, BA and BB.

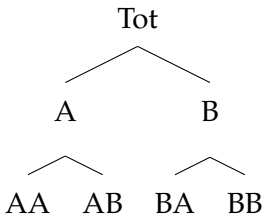

Figure 1: *Two level hierarchical diagram.*

We let $\boldsymbol{y}_t \in \mathbb{R}^n$ be a vector comprised of observations of all variables in the hierarchy at time t, and $\boldsymbol{b}_t \in \mathbb{R}^m$ is a vector comprised of observations of all bottom-level series at time t. The hierarchical structure of the data imply the following holds for all t

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

where $\boldsymbol{S}$ is an $n \times m$ constant matrix that encodes the aggregation constraints. For the hierarchy in Figure 1, $\boldsymbol{y}_t = [y_{Tot,t}, y_{A,t}, y_{B,t}, y_{C,t}, y_{AA,t}, y_{AB,t}, y_{BA,t}, y_{BB,t}]'$, $\boldsymbol{b}_t = [y_{AA,t}, y_{AB,t}, y_{BA,t}, y_{BB,t}]'$, $m = 4$, $n = 7$, and

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

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

3 Coherent forecasts

It is desirable that forecasts, whether point forecasts or probabilistic forecasts, should in some sense respect aggregation constraints. We follow other authors in using the nomenclature *coherence* to describe this property.

We now provide new definitions for coherent forecasts in terms of vector spaces that give a geometric understanding of the problem thus facilitating the development of the probabilistic forecast reconciliation in section 4.

Definition 3.1 (Coherent subspace)**.** Let an n-dimensional time series $\boldsymbol{y}_t \in \mathbb{R}^n$ be 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. The

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

We also denote the $n \times (n-m)$ orthogonal complement of $\boldsymbol{S}$ as $\boldsymbol{S}_\perp$, where $\mathfrak{s}_\perp = \text{span}\,(\boldsymbol{S}_\perp)$ is the nullspace of $\boldsymbol{S}$. Also at times is will be useful to think of pre-multiplication by $\boldsymbol{S}$ as a linear mapping from $\mathbb{R}^m$ to $\mathbb{R}^n$ in which case we use the notation $s(.)$. Although the codomain of $s(.)$ is $\mathbb{R}^n$ its image is the coherent space $\mathfrak{s}$ as depicted in Figure 2.

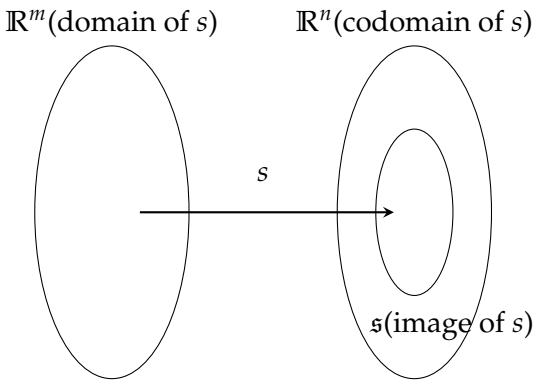

Figure 2: *The domain, codomain and image of the mapping s.*

Definition 3.2 (Coherent Point Forecasts)**.** Let $\breve{\boldsymbol{y}}_{t+h|t} \in \mathbb{R}^n$ be a point forecast of the values of all series in the hierarchy at time $t+h$, made using information up to and including time t. Then $\breve{\boldsymbol{y}}_{t+h|t}$ is *coherent* if $\breve{\boldsymbol{y}}_{t+h|t} \in \mathfrak{s}$.

Definition 3.3 (Coherent Probabilistic Forecasts)**.** Let $(\mathbb{R}^m, \mathscr{F}^{\mathbb{R}^m}, \nu)$ be a probability triple, where $\mathscr{F}^{\mathbb{R}^m}$ is the usual σ-algebra on $\mathbb{R}^m$. Let $\breve{\nu}$ be a probability measure on $\mathfrak{s}$ with σ-algebra $\mathscr{F}^{\mathfrak{s}}$. Here $\mathscr{F}^{\mathfrak{s}}$ is formed as collection of sets $s(\mathcal{B})$, where $s(\mathcal{B})$ denotes the image of the set $\mathcal{B} \in \mathscr{F}^{\mathbb{R}^m}$ under the mapping $s(.)$. Let the measure $\breve{\nu}$ have the property

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

A probabilistic forecast is coherent if uncertainty in $\boldsymbol{y}_{t+h|h}$ conditional on all information up to time t is characterised by the probability triple $(\mathfrak{s}, \mathscr{F}^{\mathfrak{s}}, \breve{\nu})$.

These definitions of the coherent space $\mathfrak{s}$ and coherent point and probabilistic forecasts are defined in terms of the mapping $s(.)$ and may give the impression that the bottom level series play an important role in the definition. However, alternative definitions could be formed using

any set of basis vectors that spans $\mathfrak{s}$. For example, consider the most simple three variable hierarchy where $y_{1,t} = y_{2,t} + y_{3,t}$. In this case the matrix $\boldsymbol{S}$ has columns $(1,1,0)'$ and $(1,0,1)'$ spanning $\mathfrak{s}$ and premultiplying by $\boldsymbol{S}$ transforms arbitrary values of $y_{2,t}$ and $y_{3,t}$ into a coherent vector for the full hierarchy. However the columns $(1,0,1)'$ and $(0,1,-1)'$ also span $\mathfrak{s}$ and define a mapping that transforms arbitrary values of $y_{1,t}$ and $y_{2,t}$ into a coherent vector for the full hierarchy. The definitions above could be made in terms of any series and not just the bottom level series. In general, we call the series (or linear combinations thereof) used in the definitions of coherence as *basis series*. Unless stated otherwise, we will always assume that the basis series are the bottom level series as in Definition 3.2 and Defintion 3.3, since this facilitates comparison with existing approaches in the literature.

To the best of our knowledge, the only other definition of coherent probabilistic forecasts is given by Ben Taieb et al. (2017) who define coherent probabilistic forecasts in terms of convolutions. According to their definition, probabilistic forecasts are coherent when a convolution of forecast distributions of disaggregate series is identical to the forecast distribution of the corresponding aggregate series. Their definition is consistent with our definition, our reason for providing a different definition is that the geometric understanding of coherence will facilitate our definitions of point and probabilistic forecast reconciliation to which we now turn our attention.

4 Forecast reconciliation

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

4.1 Point forecast reconciliation

Let $\hat{\boldsymbol{y}}_{t+h|t} \in \mathbb{R}^n$ be any set of incoherent point forecasts at time $t+h$ using information up to and including time t. Let $\boldsymbol{G}$ be an $m \times n$ matrix and $g : \mathbb{R}^n \to \mathbb{R}^m$ be the mapping corresponding to pre-multiplication by $\boldsymbol{G}$.

Definition 4.1. The point forecast $\tilde{\boldsymbol{y}}_{t+h|t}$ "reconciles" $\hat{\boldsymbol{y}}_{t+h|t}$ with respect to the mapping $g(.)$ iff

$$\tilde{\boldsymbol{y}}_{t+h|t} = \boldsymbol{SG}\hat{\boldsymbol{y}}_{t+h|t}\,.$$

In definition 4.1 $\boldsymbol{SG}$ is a projection matrix that maps $\mathbb{R}^n$ onto $\mathfrak{s}$. We note that a similar characterisation of reconciliation has been used in previous studies but where the notation $\boldsymbol{P}$ is used in

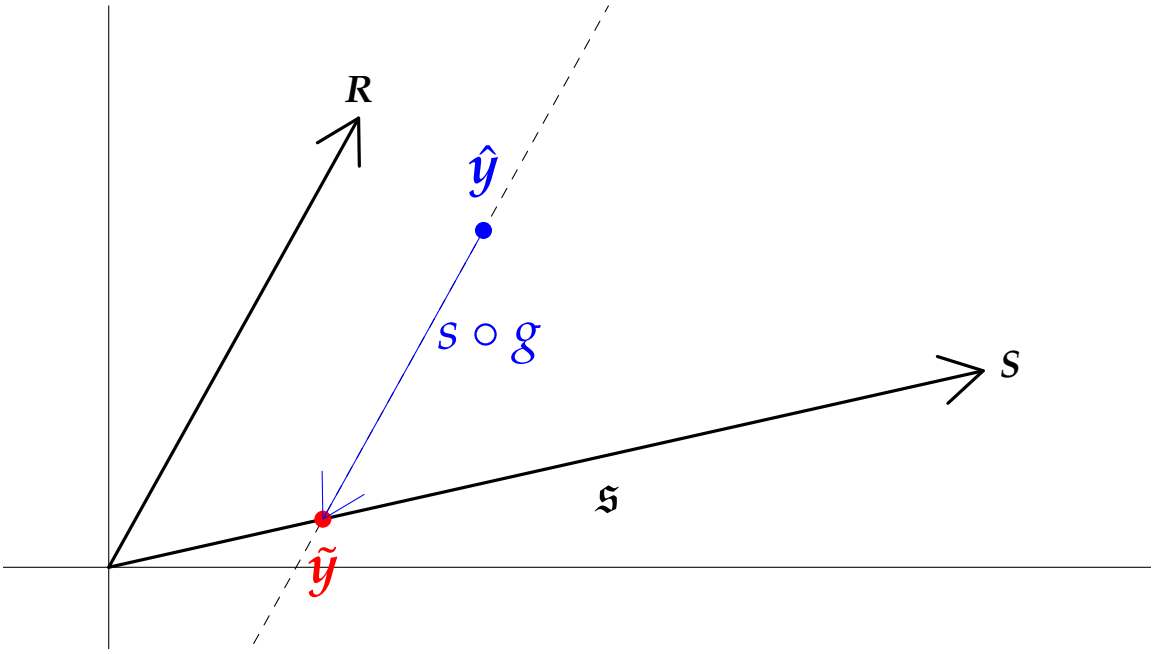

Figure 3: *Summary of probabilistic point reconciliation. The mapping $G(.)$ projects the unreconciled forecast $\hat{\boldsymbol{y}}$ onto Note that since the smallest hierarchy involves three dimensions, this figure is only a schematic*

place of $\boldsymbol{G}$. In principle the definition can be generalised where $g(.)$ is a non-linear mapping, a possibility we consider in Section **??** [AP: I WILL ADD THIS LATER].

Definition 4.1, defines an entire family of forecast reconciliation methods including many methods currently extant in the literature. To provide a geometric interpretation, we introduce an $n \times (n-m)$ matrix $\boldsymbol{R}$ whose columns span the null space $\mathfrak{s}_{\perp}$. For example, a straightforward choice of $\boldsymbol{R}$ for the most simple three variable hierarchy where $y_{1,t} = y_{2,t} + y_{3,t}$, is the vector $(1, -1, -1)$ which is orthogonal (in the Euclidean sense) to the columns of $\boldsymbol{S}$. In this case, the matrix $\boldsymbol{R}$ can be interpreted as a 'restrictions' matrix since it has the property that $\boldsymbol{R}'\boldsymbol{y} = \boldsymbol{0}$ for coherent $\boldsymbol{y}$. In general, the columns of any given $\boldsymbol{R}$ matrix can also be thought of as 'directions' along which incoherent point forecasts are projected onto the coherent space, and these directions need not be orthogonal to $\mathfrak{s}$. A schematic of this geometric interpretation is provided in Figure 3.

To illustrate further note that the columns of $\boldsymbol{S}$ and $\boldsymbol{R}$ provide a basis for $\mathbb{R}^n$. As such any incoherent set of point forecasts $\hat{\boldsymbol{y}}_{t+h|t} \in \mathbb{R}^n$, can be expressed in terms of coordinates in the basis defined by $\boldsymbol{S}$ and $\boldsymbol{R}$. Let $\tilde{\boldsymbol{b}}_{t+h|t}$ and $\tilde{\boldsymbol{a}}_{t+h|t}$ be the coordinates corresponding to $\boldsymbol{S}$ and $\boldsymbol{R}$, respectively. The process of reconciliation involves setting $\tilde{\boldsymbol{b}}_{t+h|t}$ to be the values of the reconciled bottom-level series and setting $\tilde{\boldsymbol{a}}_{t+h} = \boldsymbol{0}$ to ensure coherence. From properties of linear algebra it follows that

$$\hat{\boldsymbol{y}}_{t+h|t} = (\boldsymbol{S}\ \boldsymbol{R}) \begin{pmatrix} \tilde{\boldsymbol{b}}_{t+h|t} \\ \tilde{\boldsymbol{a}}_{t+h|t} \end{pmatrix} = \boldsymbol{S}\tilde{\boldsymbol{b}}_{t+h|t} + \boldsymbol{R}\tilde{\boldsymbol{a}}_{t+h|t},$$

while setting $\tilde{\boldsymbol{a}}_{t+h|t} = \boldsymbol{0}$ gives the reconciled point forecast

$$\tilde{\boldsymbol{y}}_{t+h|t} = \boldsymbol{S}\tilde{\boldsymbol{b}}_{t+h|t}$$

In order to find $\tilde{\boldsymbol{b}}_{t+h|t}$ we require the inverse $(\boldsymbol{S}\ \boldsymbol{R})^{-1}$ which is given by

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

where $\boldsymbol{S}_\perp$ and $\boldsymbol{R}_\perp$ be the orthogonal complements of $\boldsymbol{S}$ and $\boldsymbol{R}$ respectively. Thus it follows that $\tilde{\boldsymbol{b}}_{t+h} = (\boldsymbol{R}'_\perp \boldsymbol{S})^{-1}\boldsymbol{R}'_\perp \hat{\boldsymbol{y}}_{t+h}$ and $\tilde{\boldsymbol{y}}_{t+h|t} = \boldsymbol{S}(\boldsymbol{R}'_\perp \boldsymbol{S})^{-1}\boldsymbol{R}'_\perp \hat{\boldsymbol{y}}_{t+h|t}$. Here $(\boldsymbol{R}'_\perp \boldsymbol{S})^{-1}\boldsymbol{R}'_\perp$ is equivalent to $\boldsymbol{G}$ in Definition 4.1.

Point reconciliation methods will always minimise the distance between unreconciled and reconciled forecasts, however the specific distance will depend on the choice of $\boldsymbol{R}$. For example Hyndman et al., 2011 consider $\tilde{\boldsymbol{y}}^{OLS}_{t+h} = \boldsymbol{S}(\boldsymbol{S}'\boldsymbol{S})^{-1}\boldsymbol{S}'\hat{\boldsymbol{y}}_{t+h}$ which minimises the Euclidean distance between $\hat{\boldsymbol{y}}$ and $\tilde{\boldsymbol{y}}$. Wickramasuriya, Athanasopoulos, and Hyndman, 2018 consider $\tilde{\boldsymbol{y}}^{MinT}_{t+h} = \boldsymbol{S}(\boldsymbol{S}'\boldsymbol{W}_h^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}_h^{-1}\hat{\boldsymbol{y}}_{t+h}$, where $\boldsymbol{W}$ is an estimate of the variance covariance matrix of the unreconciled errors. This minimises the Mahalonobis distance between $\hat{\boldsymbol{y}}$ and $\tilde{\boldsymbol{y}}$. Bottom up methods minimise distance between reconciled and unreconciled forecasts only along dimensions corresponding to the bottom level series. As such bottom up methods are at should be thought of as a boundary case of reconciliation methods, since they ultimately do not use information at all levels of the hierarchy.

4.2 Probabilistic forecast reconciliation

We now extend the methodology of point forecast reconciliation to probabilistic forecasts

Let $(\mathbb{R}^n, \mathscr{F}^{\mathbb{R}^n}, \hat{\nu})$ be an probability triple, that is not necessarily coherent and that characterises forecast uncertainty for all variables in the hierarchy at time $t+h$ conditional on all information up to time t. Let $g : \mathbb{R}^n \to \mathbb{R}^m$ be a linear mapping. Let $(\mathbb{R}^m, \mathscr{F}^{\mathbb{R}^m}, \nu)$ be a probability triple defined on $\mathbb{R}^m$ where ν has the following property:

$$\nu(\mathcal{B}) = \hat{\nu}(g^{-1}(\mathcal{B})), \quad \forall \mathcal{B} \in \mathscr{F}^{\mathbb{R}^m}.$$

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

Definition 4.2. We define the reconciled probability measure of $\hat{\nu}$ with respect to the mapping $g(.)$ as a probability measure $\tilde{\nu}$ on $\mathfrak{s}$ with σ-algebra $\mathscr{F}_{\mathfrak{s}}$ where the following holds

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

This definition extends the notion of forecast reconciliation to the probabilistic setting. Under point reconciliation methods, the reconciled point forecast is equal to the unreconciled point forecast after the latter is passed through two linear mappings. Similarly, probabilistic forecast reconciliation assigns the same probability to two sets where the points in one set are obtained by passing all points in the other set through two linear mappings. This is depicted schematically in Figure 4.

We now discuss how this definition can be used in practice to obtain reconciled probabilistic forecasts for hierarchical time series. Recall that the case of forecast reconciliation could be broken down into three steps. In the first, $\hat{\boldsymbol{y}}_{t+h|t}$ is transformed into coordinates $\tilde{\boldsymbol{b}}_{t+h|t}$ and $\tilde{\boldsymbol{a}}_{t+h|t}$ via a change of basis. In the second, $\tilde{\boldsymbol{a}}_{t+h|t}$ is discarded and $\tilde{\boldsymbol{b}}_{t+h|t}$ are kept as the bottom level reconciled forecasts. In the third, reconciled forecasts for the entire hierarchy are recovered via $\tilde{\boldsymbol{y}}_{t+h|t} = \boldsymbol{S}\tilde{\boldsymbol{b}}_{t+h|t}$. We now outline these three steps for probabilistic forecasts and demonstrate how they correspond to Definition 4.2.

While $\hat{\nu}$ is a probability measure for an n-vector $\hat{\boldsymbol{y}}_{t+h|t}$, probability statements in terms of a different coordinate system can be made via an appropriate change of basis. Letting $f(.)$ be generic notation for a probability density functions and following the notation from our definition of point forecast reconciliation where $\hat{\boldsymbol{y}} = \boldsymbol{S}\tilde{\boldsymbol{b}}_{t+h|t} + \boldsymbol{R}\tilde{\boldsymbol{a}}_{t+h|t}$ we obtain

$$f(\hat{\boldsymbol{y}}_{t+h|t}) = f(\boldsymbol{S}\tilde{\boldsymbol{b}}_{t+h|t} + \boldsymbol{R}\tilde{\boldsymbol{a}}_{t+h|t})\,|\,(\boldsymbol{S}\ \boldsymbol{R})\,|$$

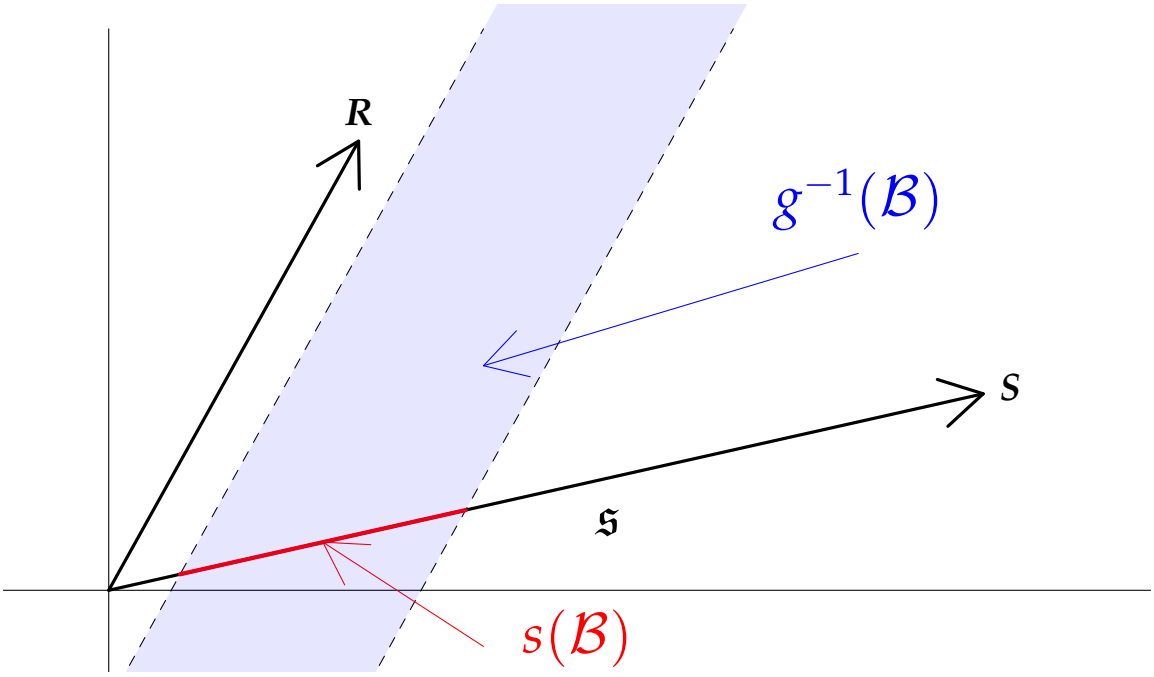

Figure 4: *Summary of probabilistic forecast reconciliation. The probability that $\boldsymbol{y}_{t+h}$ lies in the red line segment under the reconciled probabilistic forecast is defined to be equal to the probability that $\boldsymbol{y}_{t+h}$ lies in the shaded blue area under the unreconciled probabilistic forecast. Note that since the smallest hierarchy involves three dimensions, this figure is only a schematic*

The expression $\hat{v}(g^{-1}(\mathcal{B}))$ in Definition 4.2 is equivalent to the probability statement $\Pr(\hat{\boldsymbol{y}}_{t+h|t} \in g^{-1}(\mathcal{B}))$. After the change of basis this is equivalent to $\Pr(\tilde{\boldsymbol{b}} \in \mathcal{B})$ which implies

$$\begin{aligned}\Pr(\hat{\boldsymbol{y}}_{t+h|t} \in g^{-1}(\mathcal{B})) &= \int_{g^{-1}(\mathcal{B})} f(\hat{\boldsymbol{y}}_{t+h|t}) d\hat{\boldsymbol{y}}_{t+h|t} \\ &= \int_{\mathcal{B}} \int f(\boldsymbol{S}\tilde{\boldsymbol{b}}_{t+h|t} + \boldsymbol{R}\tilde{\boldsymbol{a}}_{t+h|t}) \left| (\boldsymbol{S}\ \boldsymbol{R}) \right| d\tilde{\boldsymbol{a}}_{t+h|t} d\tilde{\boldsymbol{b}}_{t+h|t}\end{aligned}$$

After integrating out over $\tilde{\boldsymbol{a}}_{t+h|t}$, a step analogous to setting $\tilde{\boldsymbol{a}}_{t+h|t} = 0$ for point forecasting, we obtain an expressions that gives the probability the reconciled bottom level series lies in the region $\mathcal{B}$. This corresponds to $v(\mathcal{B})$ in Definition 4.2. To make a valid probability statement

about the entire hierarchy we simply use the bottom level probabilistic forecasts together with Definition 3.3.

Example: Gaussian Distributions

Suppose an unreconciled probabilistic forecast is Gaussian with mean $\hat{\boldsymbol{\mu}}$ and variance-covariance matrix $\hat{\boldsymbol{\Sigma}}$. The subscripts $t+h|t$ are suppressed for brevity. The unreconciled density

$$f(\hat{\boldsymbol{y}}) = (2\pi)^{-n/2}|\hat{\boldsymbol{\Sigma}}|^{-1/2}\exp\left\{-\frac{1}{2}\left[(\hat{\boldsymbol{y}}-\hat{\boldsymbol{\mu}})'\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{y}}-\hat{\boldsymbol{\mu}})\right]\right\}$$

After a change in basis

$$f(\tilde{\boldsymbol{b}},\tilde{\boldsymbol{a}}) = (2\pi)^{-\frac{n}{2}}\left|\hat{\boldsymbol{\Sigma}}_{t+h}\right|^{-\frac{1}{2}}\left|(\boldsymbol{S}\ \boldsymbol{R})\right|\exp\left\{-\frac{1}{2}q\right\},$$

where

$$q = (\boldsymbol{S}\tilde{\boldsymbol{b}}+\boldsymbol{R}\tilde{\boldsymbol{a}}-\hat{\boldsymbol{\mu}})'\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{S}\tilde{\boldsymbol{b}}+\boldsymbol{R}\tilde{\boldsymbol{a}}-\hat{\boldsymbol{\mu}})$$

The quadratic form q can be rearranged as

$$\begin{aligned} q &= \left((\boldsymbol{S}\ \boldsymbol{R})\begin{pmatrix}\tilde{\boldsymbol{b}}\\ \tilde{\boldsymbol{a}}\end{pmatrix}-\hat{\boldsymbol{\mu}}\right)'\hat{\boldsymbol{\Sigma}}^{-1}\left((\boldsymbol{S}\ \boldsymbol{R})\begin{pmatrix}\tilde{\boldsymbol{b}}\\ \tilde{\boldsymbol{a}}\end{pmatrix}-\hat{\boldsymbol{\mu}}\right), \\ &= \left(\begin{pmatrix}\tilde{\boldsymbol{b}}\\ \tilde{\boldsymbol{a}}\end{pmatrix}-(\boldsymbol{S}\ \boldsymbol{R})^{-1}\hat{\boldsymbol{\mu}}_{t+h}\right)'\left[(\boldsymbol{SR})\hat{\boldsymbol{\Sigma}}_{t+h}(\boldsymbol{S}\ \boldsymbol{R})'\right]^{-1}\left(\begin{pmatrix}\tilde{\boldsymbol{b}}\\ \tilde{\boldsymbol{a}}\end{pmatrix}-(\boldsymbol{S}\ \boldsymbol{R})^{-1}\hat{\boldsymbol{\mu}}_{t+h}\right). \end{aligned}$$

Recall that

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

Then q can be rearranged further as

$$\begin{aligned} q &= \left[\begin{pmatrix}\tilde{\boldsymbol{b}}\\ \tilde{\boldsymbol{a}}\end{pmatrix}-\begin{pmatrix}\boldsymbol{G}\\ \boldsymbol{H}\end{pmatrix}\hat{\boldsymbol{\mu}}_{t+h}\right]'\left[\begin{pmatrix}\boldsymbol{G}\\ \boldsymbol{H}\end{pmatrix}\hat{\boldsymbol{\Sigma}}_{t+h}\begin{pmatrix}\boldsymbol{G}\\ \boldsymbol{H}\end{pmatrix}'\right]^{-1}\left[\begin{pmatrix}\tilde{\boldsymbol{b}}\\ \tilde{\boldsymbol{a}}\end{pmatrix}-\begin{pmatrix}\boldsymbol{G}\\ \boldsymbol{H}\end{pmatrix}\hat{\boldsymbol{\mu}}_{t+h}\right] \\ &= \begin{pmatrix}\tilde{\boldsymbol{b}}-\boldsymbol{G}\hat{\boldsymbol{\mu}}\\ \tilde{\boldsymbol{a}}-\boldsymbol{H}\hat{\boldsymbol{\mu}}\end{pmatrix}'\left[\begin{pmatrix}\boldsymbol{G}\\ \boldsymbol{H}\end{pmatrix}\hat{\boldsymbol{\Sigma}}_{t+h}\begin{pmatrix}\boldsymbol{G}\\ \boldsymbol{H}\end{pmatrix}'\right]^{-1}\begin{pmatrix}\tilde{\boldsymbol{b}}-\boldsymbol{G}\hat{\boldsymbol{\mu}}\\ \tilde{\boldsymbol{a}}-\boldsymbol{H}\hat{\boldsymbol{\mu}}\end{pmatrix} \end{aligned}$$

Similar manipulations on determinant of the covariance matrix lead to the following expression for the density

$$f(\tilde{\boldsymbol{b}}, \tilde{\boldsymbol{a}}) = (2\pi)^{-\frac{n}{2}} \left| \begin{pmatrix} \boldsymbol{G}\hat{\boldsymbol{\Sigma}}\boldsymbol{G}' & \boldsymbol{G}\hat{\boldsymbol{\Sigma}}\boldsymbol{H}' \\ \boldsymbol{H}\hat{\boldsymbol{\Sigma}}\boldsymbol{G}' & \boldsymbol{H}\hat{\boldsymbol{\Sigma}}\boldsymbol{H}' \end{pmatrix} \right|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \begin{pmatrix} \tilde{\boldsymbol{b}} - \boldsymbol{G}\hat{\boldsymbol{\mu}} \\ \tilde{\boldsymbol{a}} - \boldsymbol{H}\hat{\boldsymbol{\mu}} \end{pmatrix}' \begin{pmatrix} \boldsymbol{G}\hat{\boldsymbol{\Sigma}}\boldsymbol{G}' & \boldsymbol{G}\hat{\boldsymbol{\Sigma}}\boldsymbol{H}' \\ \boldsymbol{H}\hat{\boldsymbol{\Sigma}}\boldsymbol{G}' & \boldsymbol{H}\hat{\boldsymbol{\Sigma}}\boldsymbol{H}' \end{pmatrix}^{-1} \begin{pmatrix} \tilde{\boldsymbol{b}} - \boldsymbol{G}\hat{\boldsymbol{\mu}} \\ \tilde{\boldsymbol{a}} - \boldsymbol{H}\hat{\boldsymbol{\mu}} \end{pmatrix} \right\}.$$

Marginalising out $\tilde{\boldsymbol{a}}$, leads to the following bottom level reconciled forecasts.

$$f(\tilde{\boldsymbol{b}}) = (2\pi)^{-\frac{m}{2}} \left| \boldsymbol{G}\hat{\boldsymbol{\Sigma}}\boldsymbol{G}' \right|^{\frac{1}{2}} \exp\Big\{ -\frac{1}{2} (\tilde{\boldsymbol{b}} - \boldsymbol{G}\hat{\boldsymbol{\mu}})' (\boldsymbol{G}\hat{\boldsymbol{\Sigma}}\boldsymbol{G}')^{-1} (\tilde{\boldsymbol{b}} - \boldsymbol{G}\hat{\boldsymbol{\mu}}) \Big\}.$$

Which implies that the reconciled probabilistic forecast for the bottom level series is $\tilde{\boldsymbol{b}}_{t+h} \sim \mathcal{N}(\boldsymbol{G}\hat{\boldsymbol{\mu}}_{t+h}, \boldsymbol{G}\hat{\boldsymbol{\Sigma}}_{t+h}\boldsymbol{G}')$. The reconciled probabilistic forecasts for the whole hierarchy follow a degenerate Gaussian distribution with mean $\boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{\mu}}$ and rank deficient covariance matrix $\boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{\Sigma}}_{t+h}\boldsymbol{G}'\boldsymbol{S}'$.

4.3 Elliptical Distributions

We now show that if the mapping $g(.)$ in definition 4.1 is extended to allow for a translation by $\boldsymbol{d}$, i.e. $g(\boldsymbol{y}) = \boldsymbol{G}\boldsymbol{y} + \boldsymbol{d}$, then the true predictive distribution can be recovered for elliptical distributions. Here, for any square matrix $\boldsymbol{C}$, $\boldsymbol{C}^{1/2}$ and $\boldsymbol{C}^{-1/2}$ are defined to satisfy $\boldsymbol{C}^{1/2}\left(\boldsymbol{C}^{1/2}\right)' = \boldsymbol{C}$ and $\boldsymbol{C}^{-1/2}\left(\boldsymbol{C}^{-1/2}\right)' = \boldsymbol{C}^{-1}$, for example $\boldsymbol{C}^{1/2}$ may be obtained via the Cholesky or eigenvalue decompositions.

Theorem 4.1 (Reconciliation for Elliptical Distributions)**.** *Let an unreconciled probabilistic forecast come from the elliptical class with location parameter $\hat{\boldsymbol{\mu}}$ and scale matrix $\hat{\boldsymbol{\Sigma}}$. Let the true predictive distribution of $\boldsymbol{y}_{t+h|t}$ also belong to the elliptical class with location parameter $\boldsymbol{\mu}$ and scale matrix $\boldsymbol{\Sigma}$. Then the affine reconciliation mapping $g(\breve{\boldsymbol{y}}) = \boldsymbol{G}_{opt}\breve{\boldsymbol{y}} + \boldsymbol{d}_{opt}$ with $\boldsymbol{G}_{opt} = \boldsymbol{\Omega}^{1/2}\boldsymbol{\Sigma}^{-1/2}$ and $\boldsymbol{d}_{opt} = \boldsymbol{S}\boldsymbol{G}_{opt}\left(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\right)$ recovers the true predictive density, where $\boldsymbol{\Omega}$ is the true variance covariance matrix of the predictive distribution for the bottom level.*

Proof. Since elliptical distributions are closed under affine transformations, and are closed under marginalisation, reconciliation of an elliptical distribution yields an elliptical distribution

(although the unreconciled and unreconciled distributions may be different members of the class of elliptical distributions). The scale matrix of the reconciled forecast is given by $\boldsymbol{S}\boldsymbol{G}_{opt}\boldsymbol{\Sigma}\boldsymbol{G}_{opt}'\boldsymbol{S}'$ while the location matrix is given by $\boldsymbol{S}\boldsymbol{G}_{opt}\hat{\boldsymbol{\mu}} + \boldsymbol{d}_{opt}$. The reconciled scale matrix is

$$\begin{aligned}\tilde{\boldsymbol{\Sigma}}_{opt} =& \boldsymbol{S}\boldsymbol{\Omega}^{1/2}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Omega}^{1/2}\boldsymbol{S}' \\ =& \boldsymbol{S}\boldsymbol{\Omega}\boldsymbol{S}' \\ =& \boldsymbol{\Sigma}\end{aligned}$$

For the choices of $\mathbf{G}_{opt}$ and $\mathbf{d}_{opt}$ given above, the reconciled location matrix is

$$\begin{aligned}\tilde{\boldsymbol{\mu}}_{opt} =& \boldsymbol{S}\boldsymbol{G}_{opt}\hat{\boldsymbol{\mu}} + \boldsymbol{S}\boldsymbol{G}_{opt}\left(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\right) \\ =& \boldsymbol{S}\boldsymbol{G}_{opt}\hat{\boldsymbol{\mu}} - \boldsymbol{S}\boldsymbol{G}_{opt}\hat{\boldsymbol{\mu}} + \boldsymbol{S}\boldsymbol{G}_{opt}\boldsymbol{\mu} \\ =& \boldsymbol{S}\boldsymbol{G}_{opt}\boldsymbol{\mu} \\ =& \boldsymbol{\mu}\end{aligned}$$

The final equality above holds since $\boldsymbol{S}\boldsymbol{G}_{opt}$ projects any vector onto the coherent subspace, however since $\boldsymbol{\mu}$ is the mean of the true data generating process, it must already lie on the coherent subspace and $\boldsymbol{S}\boldsymbol{G}_{opt}\boldsymbol{\mu} = \boldsymbol{\mu}$. □

It should be noted that $\boldsymbol{G}_{opt}$ and $\boldsymbol{d}_{opt}$ depend on the true location and scale of the predictive distribution and are thus infeasible in practice. However, these expressions do provide some insight on why some choices for $\boldsymbol{G}$ in the point forecast reconciliation literature may also work well for probabilistic forecasts. To illustrate, first rearrange the equation for the optimal value of $\boldsymbol{G}_{opt} = \boldsymbol{\Omega}^{1/2}\hat{\boldsymbol{\Sigma}}^{-1/2}$ as $\boldsymbol{\Omega}^{1/2} = \boldsymbol{G}_{opt}\hat{\boldsymbol{\Sigma}}^{1/2}$. For arbitrary $\boldsymbol{G}\hat{\boldsymbol{\Sigma}}^{1/2}$ is an approximation for $\boldsymbol{\Omega}^{1/2}$. This approximation has similarities with the way bottom level estimates are produced for point forecast reconciliation. Rather than mapping a vector of point forecasts to reconcilied bottom level forecasts, the columns of $\hat{\boldsymbol{\Sigma}}^{1/2}$ are mapped to an approximation of the columns of $\boldsymbol{\Omega}^{1/2}$. The value for $\tilde{\boldsymbol{\mu}}_{opt}$ includes a projection of $\hat{\boldsymbol{\mu}}$ onto $\mathfrak{s}$. While it is infeasible to obtain the translation $\boldsymbol{d}$, it is worthwhile noting that this measures the difference between the reconciled mean and true mean.

5 Evaluation of hierarchical probabilistic forecasts

The necessary final step in hierarchical forecasting is to make sure that our forecast distributions are accurate. In general, forecasters prefer to maximize the sharpness of the forecast distribution subject to 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 predictive 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 follow the latter convention here.

Let $\boldsymbol{P}$ be a forecast distribution and let $\boldsymbol{Q}$ be the true data generating process respectively. Further let ω be a realization from $\boldsymbol{Q}$. Then a scoring rule is a function that maps P, ω to $\mathbb{R}$. It is a "proper" scoring rule if

$$\mathrm{E}_{\boldsymbol{Q}}[S(P,\omega)] \leq \mathrm{E}_{\boldsymbol{Q}}[S(Q,\omega)], \tag{1}$$

where $\mathrm{E}_{\boldsymbol{Q}}[S(\boldsymbol{P},\boldsymbol{\omega})]$ is the expected score under the true distribution $\boldsymbol{Q}$ (Gneiting et al., 2008; Gneiting and Katzfuss, 2014). When this inequality is strict, the scoring rule is said to be strictly proper.

In the context of probabilistic forecast reconciliation there could be two motivations for using scoring rules. The first is to compare unreconciled densities to reconciled densities. Although reconciliation is a valuable goal in and of itself since it can be important in aligning decision making across, for example, different units of an enterprise, in the point forecasting literature, forecast reconciliation has also been shown to improve forecast performance . It will be worthwhile to see whether the same holds in the probabilistic forecasting case . The second motivation

Some References

mention energy score proof

for using scoring rules is to compare two or more sets of reconciled probabilistic forecasts. The objective here is to evaluate which reconciliation mapping $g(.)$ works best in practice.

5.1 Univariate Scoring rules

One way to evaluate probabilistic forecasts is via the application of univariate scoring rules to each variable in a hierarchy. A summary can be taken of the expected scores across each margin for example a mean or median. In the simulations of section 6 we consider two scoring rules. The log score is given by the log of the marginal density of each variable. The cumulative rank probability score generalises mean square error and is given by

$$\begin{aligned}\text{CRPS}(\breve{F}_i, y_{T+h,i}) &= \int \left(F_i(\breve{y}_i) - \mathbb{1}(y_i < y_{T+h,i})\right) dy_i \\ &= \mathrm{E}_{\breve{F}_i}|\breve{Y}_{T+h,i} - y_{T+h,i}| - \frac{1}{2}\mathrm{E}_{\breve{F}_i}|\breve{Y}_{T+h,i} - \breve{Y}^*_{T+h,i}|\,,\end{aligned}$$

where the expectations in the second line can be approximated by Monte Carlo.

Although univariate scoring ruled have been used in the limited literature on probabilistic forecasting for hierarchies Ben Taieb et al., 2017 and Jeon et al to date, there are a number of shortcomings to this approach. Crucially, evaluating univariate scores on the margins do not account for the dependence in the hierarchy.

Jeon Panagiotelis and Petropoulos reference

5.2 Multivariate Scoring rules

While the a number of alternative proper scoring rules are available for univariate forecasts, the multivariate case is somewhat more limited. Here we focus on three cases, the log score, the energy score and the variogram score. These are summarised in table 1 summarizes a few existing proper scoring rules.

The log score can be approximated using a sample of values from the probabilistic forecast density (Jordan, Krüger, and Lerch, 2017) however it is more commonly used when a parametric form for the density is available for the probabilistic forecast. So far, we have only defined probabilistic forecasts in terms of probability measures. Although densities can be obtained for both reconciled and unreconciled forecasts, the degeneracy of reconciled forecasts is problematic when using log scores. We will discuss this further in the next subsection.

The energy score on the other hand can be defined in terms of the characteristic function of the probabilistic forecast, but the representation in Table 1 in terms of expectations leads itself to

easy computation when samples from the probabilistic forecast are available. An interesting case is where $\alpha = 2$, where it can be easily shown that

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

where $\breve{\boldsymbol{\mu}}_{T+h} = \mathrm{E}_{\breve{\boldsymbol{F}}}(\breve{\boldsymbol{Y}}_{T+h})$. In this limiting case, the energy score only measures the accuracy of the forecast mean, and not the entire distribution and the energy score is proper, but not strictly proper. Pinson and Tastu (2013) also argues that the energy score has very low discriminative 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 $p = 0.5$.

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

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

5.2.1 Comparing Unreconciled Forecasts to Reconciled Forecasts

For both reconciled and unreconciled densities it is possible to obtain a density from the probability measures defined in 3. As such it may seem sensible to compare unreconciled densities to reconciled densities on the basis of log score. However more careful consideration shows that using the log score may fail in the case of multivariate distributions with a degeneracy.

Consider a rotated version of hierarchical time series $z_t = \boldsymbol{U}\boldsymbol{y}_t$ so that the first m elements of z_t denoted $z(1)_t$ are unconstrained, while the remaining $n-m$ elements denoted $z(2)_t$ equal 0 when the aggregation constraints hold. An example of the $n \times n$ $\boldsymbol{U}$ could be the left singular values of the matrix $\boldsymbol{S}$. For a non-degenerate probability measure on $\mathbb{R}^n$, the density is the Radon-Nikodym derivative with respect to the usual Lesbegue measure on $\mathbb{R}^n$. For a coherent probability measure the density after rotation is the Radon-Nikodym derivative with respect to the usual Lesbegue measure on $\mathbb{R}^m$.

Consider the case where the true density is $f_1(z_t^{(1)})\mathbb{1}\left(z_t^{(2)}\right)$, the reconciled density is identical to the true density and the unreconciled density is given by $f_1(z_t^{(1)})f_2(z_t^{(2)})$, where f_2 is highly concentrated around 0 but still a proper density, for example a Gaussian with variance $\sigma^2\boldsymbol{I}$ with $\sigma^2 < (2\pi)^{-1}$. The log score of the reconciled density is

$$S\left(\tilde{f}, z_t^{(1)}\right) = -\log f_1\left(z_t^{(1)}\right),$$

while that of the unreconciled density is

$$\begin{aligned} S\left(\hat{f}, z_t^{(1)}\right) &= -\log f_1(z_t^{(1)}) - f_2(z_t^{(1)}) \\ &= -\log f_1(z_t^{(1)}) + \frac{n-m}{2}\log(2\pi\sigma^2) \\ &\leq -\log f_1(z_t^{(1)}). \end{aligned}$$

After taking expectations $ES(f, f) > ES(\hat{f}, f)$, violating the condition (1) for a proper scoring rule. A similar issue also arises when discrete random variables are modelled as if they were continuous, an issue discussed in Section 4.1, page 366 of Gneiting and Raftery, 2007. This implies that the log score should not be used to evaluate multivariate densities with degeneracies and should be avoided when comparing reconciled and unreconciled probabilistic forecasts.

Stuff below still open

To be discussed: Energy score - still trying to prove that reconciled dominates unreconciled, will show "half" proof if necessary. Variogram score - it might be possible to show reconciled dominate unreconciled, I will have a look.

5.2.2 Comparing Reconciled Forecasts to one another

A characteristic of coherent probabilistic forecasts is that they can be completely characterised in terms of basis series. If probabilistic forecasts of the basis series are accurate, then forecasts for the whole hierarchy are also accurate. Therefore, we propose that two different sets of reconciled forecasts can be compared to one another 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.

Two questions below

- Can we prove that if reconciled forecast A is better than forecast B in terms of the bottom level then it will also be better for the whole series. This is kind of obvious for log score, but may not be for energy score and variogram score
- Does the basis we use matter. We know that energy score is invariant to orthogonal rotations but S is not (semi) orthogonal. On the other hand, I don't think variogram is invariant to rotations as all unless $p = 2$.

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, $\mathrm{E}[Y_{T+h,i} \mid y_{1,i}, \ldots, y_{T,i}]$, $i = 1, \ldots, n$. Suppose we independently fit time series models for each series in the hierarchy. Then the point forecasts, $\hat{Y}_{T+h,i}$, from the estimated models are unbiased and consistent estimators of $\mathrm{E}[Y_{T+h,i} \mid y_{1,i}, \ldots, y_{T,i}]$, 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,

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

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

$$\hat{Y}_{T+h,i} \xrightarrow{p} \mathrm{E}[Y_{T+h,i} \mid y_{1,i}, \ldots, y_{T,i}] \quad \text{as} \quad T \to \infty. \tag{3}$$

Let $\hat{\boldsymbol{Y}}_{T+h} = (\hat{Y}_{T+h,1}, \ldots, \hat{Y}_{T+h,n})'$ and suppose (3) holds for $i = 1, \ldots, n$. Then from Slutsky's theorem it follows that

$$\hat{\boldsymbol{Y}}_{T+h} \xrightarrow{p} \mathrm{E}[\boldsymbol{Y}_{T+h} | \boldsymbol{\mathcal{I}}_T] \quad \text{as} \quad T \to \infty. \tag{4}$$

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

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

From (4) 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

$$\widehat{\boldsymbol{Y}_{T+h,i}|\boldsymbol{\mathcal{I}}_T} \sim \mathcal{N}(\hat{\boldsymbol{Y}}_{T+h}, \boldsymbol{W}_{T+h}) \tag{5}$$

Since our primary objective is to find the coherent forecast density of the hierarchy, we need to reconcile (5). Using (**??**), 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 (2), 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 $\boldsymbol{R}_\perp'=\boldsymbol{S}'\boldsymbol{W}_{T+h}^{-1}$.

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.

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)	$\hat{\boldsymbol{W}}^{wls}_{T+1} = \text{Diag}(\hat{\boldsymbol{W}}^{shr}_{T+1})$	$\boldsymbol{S}'(\hat{\boldsymbol{W}}^{wls}_{T+1})^{-1}$

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

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

where $u_t \sim \mathcal{N}(0, \sigma_u^2)$ 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.

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

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	

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.