Probabilistic Forecasts in Hierarchical Time Series

Doctor of Philosophy

Mid Candidature Review

Candidate:
Puwasala Gamakumara

Supervisors:
Professor Rob J Hyndman
Associate Professor George Athanasopoulos
Dr Anastasios Panagiotelis

Department of Econometrics and Business Statistics
Monash University
March, 2018

Overview of the thesis

Forecasting hierarchical time series has been of great interest in many applications. In hierarchical time series, it is important to have "coherent" forecasts across the hierarchy. That is aggregates of the forecasts at lower levels should be equal to the forecasts at the upper levels of aggregation. While there is a rich literature on coherent hierarchical point forecasting, this research focuses on a probabilistic framework for coherent hierarchical forecasts through reconciliation. Structure of the research can be outlined as follows.

Chapter 1: Introduction

- Background and motivation to the study
- Objectives
- Outline of the succeeding chapters

Chapter 2: Literature review

This chapter contains a thorough literature review of existing hierarchical point forecasting methods. Further, we discuss the importance of probabilistic framework in time series. Also, this chapter will review existing scoring rules for evaluating multivariate probabilistic forecasts in time series. We are planning to include this in a book chapter series in "Advanced Studies in Theoretical and Applied Econometrics".

Chapter 3: Probabilistic forecast reconciliation of hierarchical time series

The chapter will be initiated by providing new definitions for coherent point forecasts, coherent probabilistic forecasts and point forecast reconciliations in the hierarchical framework. This is followed by a rigorous definition for probabilistic forecast reconciliation. It will also discuss the importance of reconciliation which is followed by the accuracy measures for probabilistic hierarchical forecasts. Further, this chapter provides theoretical results on probabilistic forecast reconciliation in the Gaussian framework. These results will be evaluated via an extensive simulation study. We have a working paper on this chapter which we are planning to send to Journal of American Statistical Association (JASA).

Chapter 4: Non-parametric bootstrap approach for probabilistic hierarchical forecast reconciliation

This chapter introduces a novel non-parametric approach based on bootstrapped training errors, for obtaining coherent probabilistic forecasts. This involves producing future paths of the hierarchy by implicitly modeling the dependency structure of the hierarchy via bootstrapped training errors. Then these future paths will be reconciled to obtain coherent probabilistic forecasts. This method will be further evaluated via an extensive simulation study. We are planning to write a paper on this chapter and submit to the journal of Computational Statistics and Data Analysis.

Chapter 5: Application

- Probabilistic forecasts of domestic tourism flow in Australia.

  Domestic tourism flow in Australia can be formed in a natural geographical hierarchy. Point forecasts reconciliation of these were extensively studied in literature where the focus on this chapter is to extend that into the probabilistic framework. The non-parametric approach introduced in Chapter 4 will be applied in producing coherent probabilistic forecasts of domestic tourism flow in Australia.

- Probabilistic forecasts of Walmart sales.

Chapter 6: Conclusions

Table 1: *Time plan of the Research*

Thesis Chapter	Task description	Time duration	Progress
Introduction and Literature Review	Understanding the problem and review on literature associate with the problem.	Feb/2016 - Nov/2016	Completed
	Writting the chapters.	May/2018 - July/2018	Incomplete
Probabilistic forecast reconciliation in the hierarchical time series	Defining Coherent forecasts.	Mar/2017 - June/2017	Completed
	Defining probabilistic forecast reconciliation.	Mar/2017 - June/2017	Completed
	Gaussian forecast reconciliation.	Nov/2016 - Feb/2017	Completed
	Simulation study.	Feb/2017 - June/2017	Completed
Probabilistic forecast reconciliation in the non-parametric framework	Methodology	Nov/2016 - Feb/2017	Completed
	Simulation study	Mar/2017 - June/2017	Completed
	Providing theoretical foundation	Apr/2018 - July/2018	Incomplete
Application	Forecasting Australian Domestic Tourism Flow	Dec/2017 - Mar/2017	Complete
	Forecasting Walmart sales	Oct/2017 - Mar/2019	Incomplete

Chapter 3

Probabilistic forecast reconciliation in hierarchical time series

3.1 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 (Fliedner, 2001; Kahn, 1998; Lapide, 1998; Schwarzkopf, Tersine, and Morris, 1988). 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 (Erven and Cugliari, 2014; Hyndman et al., 2011; 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 predictive distribution is first generated, and then aggregated to obtain coherent probabilistic forecasts of the upper levels of the hierarchy. Hence this method is a bottom-up approach. They propose to first use the MinT algorithm to reconcile the means of the bottom level forecast distributions, and then a copula-based approach is employed to model the dependency structure of the hierarchy. The resulting multi-dimensional distribution is used to generating empirical forecast distributions for all bottom-level series. Thus, while Ben Taieb et al. (2017) provide coherent probabilistic forecasts, they do no forecast reconciliation of the distributions. In that sense, their approach is analogous to bottom-up point forecasting rather than forecast reconciliation.

After introducing our notation in Section 3.2, we define what is meant by probabilistic forecast reconciliation for hierarchical time series in Section 3.3. 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 3.4, we discuss in detail how this degeneracy will be taken care of in probabilistic forecast reconciliation, and in Section 3.5 we consider the evaluation of probabilistic hierarchical forecasts.

Some theoretical results on probabilistic forecast reconciliation in the Gaussian framework are given in Section 3.6, 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 3.7.

3.2 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

$$\boldsymbol{y}_t = \boldsymbol{S}\boldsymbol{b}_t, \tag{3.2.1}$$

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 3.1.

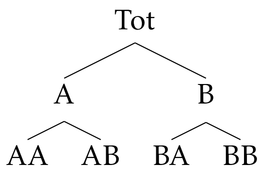

Figure 3.1: *Two level hierarchical diagram*

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

$$\boldsymbol{S} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ & \multicolumn{2}{c}{I_4} & \end{pmatrix},$$

where I_4 is a 4-dimension identity matrix.

3.3 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 3.3.1 (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 3.3.2 (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 $\breve{\boldsymbol{y}}_{t+h|t} \in \mathbb{C}^m$.*

Definition 3.3.3 (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, $(\mathbb{C}^m, \mathscr{F}_S, \breve{\nu})$ is said to be a coherent probability triple 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 3.3.3 in Figure 3.2, showing $\boldsymbol{S}$ mapping any set $\boldsymbol{A} \in \mathbb{R}^m$ to the coherent space.

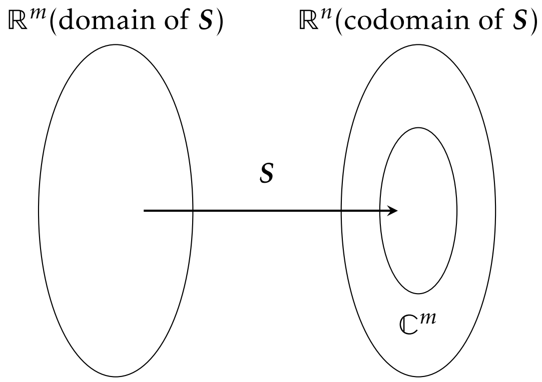

Figure 3.2: *Any set $\boldsymbol{A} \in \mathbb{R}^m$ will be mapped to $\mathbb{C}^m$ through the mapping $\boldsymbol{S}$*

Definition 3.3.3 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 3.3.3 is not unique, and one can redefine the coherent probabilistic forecasts with respect to any basis. However, we stick to Definition 3.3.3 and consider the basis defined by the columns of $\boldsymbol{S}$ in what follows.

Definitions 3.3.2 and 3.3.3 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.

3.4 Forecast reconciliation

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

3.4.1 Point forecast reconciliation

Definition 3.4.1 *Let $\hat{\boldsymbol{y}}_{t+h} \in \mathbb{R}^n$ be any set of incoherent forecasts at time $t+h$, and let*

$$\tilde{\boldsymbol{y}}_{t+h} = \boldsymbol{S} \circ \boldsymbol{g}(\hat{\boldsymbol{y}}_{t+h}), \tag{3.4.1}$$

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 3.4.1 allows for both linear and non-linear reconciliation. In other words, if $\boldsymbol{g}$ is a non-linear function, then the reconciliation of $\hat{\boldsymbol{y}}_{t+h}$ will be non-linear, while if $\boldsymbol{g}$ is a linear function, then $\boldsymbol{S} \circ \boldsymbol{g}(\cdot)$ will linearly project incoherent point forecasts onto $\mathbb{C}^m$. Previous studies in hierarchical point forecasting have only focussed on the linear case, $\boldsymbol{g}(\hat{\boldsymbol{y}}) = \boldsymbol{P}\hat{\boldsymbol{y}}$, 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 3.4.1, 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 3.4.1, 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 corresponds to the coordinates of the basis $\{\boldsymbol{s}_1, \dots, \boldsymbol{s}_m\}$. Similarly, $\tilde{\boldsymbol{t}}_{t+h}$ is another basis series corresponding to the coordinates 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

$$(\tilde{\boldsymbol{b}}'_{t+h},\ \tilde{\boldsymbol{t}}'_{t+h})' = (\boldsymbol{S} \vdots \boldsymbol{R})^{-1}\hat{\boldsymbol{y}}_{t+h}. \tag{3.4.2}$$

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,

$$(\boldsymbol{S} \vdots \boldsymbol{R})^{-1} = \begin{pmatrix} (\boldsymbol{R}'_{\perp}\boldsymbol{S})^{-1}\boldsymbol{R}'_{\perp} \\ \cdots \\ (\boldsymbol{S}'_{\perp}\boldsymbol{R})^{-1}\boldsymbol{S}'_{\perp} \end{pmatrix}. \tag{3.4.3}$$

Thus we have,

$$\begin{pmatrix} \tilde{\boldsymbol{b}}_{t+h} \\ \cdots \\ \tilde{\boldsymbol{t}}_{t+h} \end{pmatrix} = \begin{pmatrix} (\boldsymbol{R}'_{\perp}\boldsymbol{S})^{-1}\boldsymbol{R}'_{\perp} \\ \cdots \\ (\boldsymbol{S}'_{\perp}\boldsymbol{R})^{-1}\boldsymbol{S}'_{\perp} \end{pmatrix}\hat{\boldsymbol{y}}_{t+h}. \tag{3.4.4}$$

From (3.4.4) it follows that,

$$\tilde{\boldsymbol{b}}_{t+h} = (\boldsymbol{R}'_{\perp}\boldsymbol{S})^{-1}\boldsymbol{R}'_{\perp}\hat{\boldsymbol{y}}_{t+h} \tag{3.4.5}$$

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,

$$\tilde{\boldsymbol{y}}_{t+h} = \boldsymbol{S}(\boldsymbol{R}'_{\perp}\boldsymbol{S})^{-1}\boldsymbol{R}'_{\perp}\hat{\boldsymbol{y}}_{t+h}, \qquad \tilde{\boldsymbol{y}}_{t+h} \in \mathbb{C}^m < \mathbb{R}^n. \tag{3.4.6}$$

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,

$$\tilde{\boldsymbol{y}}^{OLS}_{t+h} = \boldsymbol{S}(\boldsymbol{S}'\boldsymbol{S})^{-1}\boldsymbol{S}'\hat{\boldsymbol{y}}_{t+h}. \tag{3.4.7}$$

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

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

and thus,

$$\tilde{\boldsymbol{y}}^{MinT}_{t+h} = \boldsymbol{S}(\boldsymbol{S}'\boldsymbol{W}_h^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}_h^{-1}\hat{\boldsymbol{y}}_{t+h}. \tag{3.4.8}$$

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

3.4.2 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 3.4.2 *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*

$$\nu^m(\boldsymbol{A}) = \hat{\nu}(\boldsymbol{g}^{-1}(\boldsymbol{A})), \quad \forall \quad \boldsymbol{A} \in \mathscr{F}^m. \tag{3.4.9}$$

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

$$\tilde{\nu}(\boldsymbol{S}(\boldsymbol{A})) = \hat{\nu}(\boldsymbol{g}^{-1}(\boldsymbol{A})) \qquad \forall \boldsymbol{A} \in \mathscr{F}^m, \tag{3.4.10}$$

where $\boldsymbol{S} : \mathbb{R}^m \to \mathbb{C}^m$*,* $\tilde{\nu}(\cdot)$ *is the probability measure on the measure space* $(\mathbb{C}^m, \mathscr{F}_S)$ *and* $\boldsymbol{g}^{-1}(\boldsymbol{A})$ *is the pre-image of* $\boldsymbol{A}$ *in* $\mathbb{R}^n$

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 (3.4.2). Suppose $\hat{f}(\cdot)$ is the probability density of $\hat{\boldsymbol{y}}_{t+h}$. Our goal is to reconcile $\hat{f}(\cdot)$ such that the density lives on $\mathbb{C}^m$. In order to obtain this reconciled density, we need to project $\hat{f}(\hat{\boldsymbol{y}}_{t+h})$ 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 (3.4.2), and standard results for densities of transformed variables, that

$$f_{\boldsymbol{B}}(\tilde{\boldsymbol{b}}_{t+h}, \tilde{\boldsymbol{t}}_{t+h}) = \hat{f}(\boldsymbol{S}\tilde{\boldsymbol{b}}_{t+h} + \boldsymbol{R}\tilde{\boldsymbol{t}}_{t+h}) \left| \boldsymbol{S} \vdots \boldsymbol{R} \right|, \tag{3.4.11}$$

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 (3.4.11) over the range of $\tilde{\boldsymbol{t}}_{t+h}$. This will result in the reconciled density of the bottom level series $\tilde{\boldsymbol{b}}_{t+h}$,

$$\tilde{f}(\tilde{\boldsymbol{b}}_{t+h}) = \int_{\lim(\tilde{\boldsymbol{t}}_{t+h})} \hat{f}(\boldsymbol{S}\tilde{\boldsymbol{b}}_{t+h} + \boldsymbol{R}\tilde{\boldsymbol{t}}_{t+h}) \left| \boldsymbol{S} \vdots \boldsymbol{R} \right| d\tilde{\boldsymbol{t}}_{t+h}. \tag{3.4.12}$$

Finally to get the reconciled density of the whole hierarchy, we simply follow Definition 3.3.3 to obtain

$$\tilde{f}(\tilde{\boldsymbol{y}}_{t+h}) = \boldsymbol{S} \circ \tilde{f}(\tilde{\boldsymbol{b}}_{t+h}). \tag{3.4.13}$$

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 $\mathcal{N}(\hat{\boldsymbol{\mu}}_{t+h}, \hat{\Sigma}_{t+h}) \overset{d}{\leftrightarrow} \hat{f}(\hat{\boldsymbol{y}}_{t+h})$ is an incoherent forecast distribution at time $t+h$. Then from (3.4.11) 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

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

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

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

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

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

This is the joint multivariate Gaussian distribution of $(\tilde{\boldsymbol{b}}'_{t+h} \vdots \tilde{\boldsymbol{t}}'_{t+h})'$. Then from (3.4.12) and the properties of the multivariate Gaussian distribution, it follows that

$$\tilde{f}(\tilde{\boldsymbol{b}}_{t+h}) = \frac{1}{(2\pi)^{\frac{n}{2}}\left|\boldsymbol{P}\hat{\Sigma}_{\boldsymbol{t+h}}\boldsymbol{P}'\right|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\tilde{\boldsymbol{b}}_{t+h} - \boldsymbol{P}\hat{\boldsymbol{\mu}}_{t+h})'(\boldsymbol{P}\hat{\Sigma}_{\boldsymbol{t+h}}\boldsymbol{P}')^{-1}(\tilde{\boldsymbol{b}}_{t+h} - \boldsymbol{P}\hat{\boldsymbol{\mu}}_{t+h})\right\}. \tag{3.4.14}$$

Equation (3.4.14) implies $\tilde{\boldsymbol{b}}_{t+h} \sim \mathcal{N}(\boldsymbol{P}\hat{\boldsymbol{\mu}}_{t+h}, \boldsymbol{P}\hat{\Sigma}_{t+h}\boldsymbol{P}')$ where $\boldsymbol{P} = (\boldsymbol{R}'_{\perp}\boldsymbol{S})^{-1}\boldsymbol{R}'_{\perp}$. Then from (3.4.13) it follows that

$$\tilde{f}(\tilde{\boldsymbol{y}}_{t+h}) = \boldsymbol{S} \circ \tilde{f}(\tilde{\boldsymbol{b}}_{t+h}) = \tilde{f}(\boldsymbol{S}\tilde{\boldsymbol{b}}_{t+h}). \tag{3.4.15}$$

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

3.5 Evaluation of hierarchical probabilistic forecasts

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

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

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

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

$$\mathrm{E}_{\boldsymbol{G}}[S(\boldsymbol{Y}, \boldsymbol{y})] \leq \mathrm{E}_{\boldsymbol{G}}[S(\check{\boldsymbol{Y}}, \boldsymbol{y})], \tag{3.5.1}$$

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

Table 3.1 summarizes few existing proper scoring rules.

Even though the log score can be used evaluate simulated forecast densities with large samples (Jordan, Krüger, and Lerch, 2017), it is more convenient to use if it is reasonable to assume a parametric forecast density for the hierarchy. However, the "degenerecy" of coherent forecast densities would be problematic when using log scores.

To see this, suppose the true density is degenerate at $x = 0$, i.e. $f(x) = \mathbb{1}\{x = 0\}$. Now consider two predictive densities $p_1(x)$ and $p_2(x)$. Let $p_1(x)$ is equivalent to the true density, i.e. $p_1(x) = \mathbb{1}\{x = 0\}$ and $p_2(x) =^d N(0,\sigma^2)$ with $\sigma^2 < (2\pi)^{-1}$. The expected log score of p_1 is:

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

and that of p_2 is:

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

Therefore $S(f,f) > S(p_2,f)$ and hence there exist at least one forecast density which breaks the condition (3.5.1) for proper scoring rule. This implies log score will be problematic in evaluating degenerate densities.

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

$$\mathrm{eS}(\boldsymbol{Y}_{T+h}, \breve{y}_{T+h}) = ||\boldsymbol{y}_{T+h} - \breve{\boldsymbol{\mu}}_{T+h}||^2, \tag{3.5.2}$$

where $\breve{\boldsymbol{\mu}}_{T+h} = \mathrm{E}_{\boldsymbol{F}}(\breve{\boldsymbol{Y}}_{T+h})$. Therefore in the limiting case, the energy score only measures the accuracy of the forecast mean, but not the entire distribution. Further Pinson and Tastu (2013) argued that the Energy score given in Table 3.1 has a very low discrimination ability for incorrectly specified covariances, even though it discriminates the misspecified means well.

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

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

$$\mathrm{VS}(\breve{\boldsymbol{F}}, \boldsymbol{y}_{T+h}) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(|y_{T+h,i} - y_{T+h,j}|^p - \frac{1}{B}\sum_{k=1}^{B}|\breve{Y}^k_{T+h,i} - \breve{Y}^k_{T+h,j}|^p\right)^2. \tag{3.5.3}$$

Table 3.1: *Scoring rules to evaluate multivariate forecast densities. $\breve{\boldsymbol{y}}_{T+h}$ and $\breve{\boldsymbol{y}}^*_{T+h}$ be two independent random vectors from the coherent forecast distribution $\breve{\boldsymbol{F}}$ with the density function $\breve{\boldsymbol{f}}(\cdot)$ 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 ith and jth components of the vector $\breve{\boldsymbol{Y}}_{T+h}$. Further the variogram score is given for order p where, w_{ij} are non-negative weights.*

Scoring rule	Expression	Reference
Log score	$\mathrm{LS}(\breve{\boldsymbol{F}}, \boldsymbol{y}_{T+h}) = -\log \breve{\boldsymbol{f}}(\boldsymbol{y}_{T+h})$	Gneiting and Raftery (2007)
Energy score	$\mathrm{eS}(\breve{\boldsymbol{Y}}_{T+h}, \boldsymbol{y}_{T+h}) = \mathrm{E}_{\breve{\boldsymbol{F}}}\|\|\breve{\boldsymbol{Y}}_{T+h} - \boldsymbol{y}_{T+h}\|\|^\alpha - \frac{1}{2}\mathrm{E}_{\breve{\boldsymbol{F}}}\|\|\breve{\boldsymbol{Y}}_{T+h} - \breve{\boldsymbol{Y}}^*_{T+h}\|\|^\alpha, \quad \alpha \in (0,2]$	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)

3.5.1 Evaluating coherent forecast densities

As it was mentioned in the previous section, any coherent hierarchical forecast density is a degenerate density. To the best of our knowledge, there is no proper multivariate scoring rule in literature to evaluate degenerate densities. Further, we show that some of the existing scoring rules breakdown under the degeneracy.

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

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

3.5.2 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. This can be done by evaluating the individual margins of the forecast density, using univariate proper scoring rules. Most widely used Continuous Ranked Probability Score (CRPS) would be helpful for this.

$$\mathrm{CRPS}(\breve{F}_i, y_{T+h,i}) = \mathrm{E}_{\breve{F}_i}|\breve{Y}_{T+h,i} - y_{T+h,i}| - \frac{1}{2}\mathrm{E}_{\breve{F}_i}|\breve{Y}_{T+h,i} - \breve{Y}^*_{T+h,i}|, \tag{3.5.4}$$

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

3.6 Probabilistic forecast reconciliation in the Gaussian framework

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

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

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

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

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

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

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

$$\hat{Y}_{T+h,i} \xrightarrow{p} \mathrm{E}[Y_{T+h,i} \mid y_{1,i}, \ldots, y_{T,i}] \quad \text{as} \quad T \to \infty. \tag{3.6.1}$$

Let $\hat{\boldsymbol{Y}}_{T+h} = (\hat{Y}_{T+h,1}, \ldots, \hat{Y}_{T+h,n})'$ and (3.6.1) holds for all $i = 1, \ldots, n$. Then from Slutsky's theorem it follows that,

$$\hat{\boldsymbol{Y}}_{T+h} \xrightarrow{p} \mathrm{E}[\boldsymbol{Y}_{T+h} | \boldsymbol{\mathcal{I}}_T] \quad \text{as} \quad T \to \infty. \tag{3.6.2}$$

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

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

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

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

From (3.6.2) it immediately follows that,

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

$$\boldsymbol{W}_{T+h} \xrightarrow{p} \mathrm{Var}(\boldsymbol{Y}_{T+h} | \boldsymbol{\mathcal{I}}_T) \quad \text{as} \quad T \to \infty, \tag{3.6.3}$$

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

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

$$\widehat{\boldsymbol{Y}_{T+h,i} | \boldsymbol{\mathcal{I}}_T} \sim \mathcal{N}(\hat{\boldsymbol{Y}}_{T+h}, \boldsymbol{W}_{T+h}) \tag{3.6.4}$$

Table 3.2: *Summarizing different estimates of $\boldsymbol{R}'_{\perp}$. For $n < T$, $\hat{\boldsymbol{W}}^{sam}_{T+1}$ is an unbiased and consistent estimator for $\boldsymbol{W}_{T+1}$. $\hat{\boldsymbol{W}}^{shr}_{T+1}$ is a shrinkage estimator which is much suitable for large dimensions. $\hat{\boldsymbol{W}}^{shr}_{T+1}$ was proposed by Schäfer and Strimmer (2005) and also used by Wickramasuriya, Athanasopoulos, and Hyndman (2018), where $\tau = \frac{\sum_{i\neq j}\hat{Var}(\hat{r}_{ij})}{\sum_{i\neq j}\hat{r}^2_{ij}}$, $\hat{r}_{ij}$ is the ijth element of sample correlation matrix. Further Diag($\boldsymbol{A}$) denote the diagonal matrix of $\boldsymbol{A}$*

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

Since our primary objective is to find the coherent forecast density of the hierarchy, we need to reconciled (3.6.4). Recalling from example 2, the reconciled Gaussian predictive distribution is then given by,

$$\widetilde{\boldsymbol{Y}_{T+h,i}|\mathcal{I}_T} \sim \mathcal{N}(\boldsymbol{S}\boldsymbol{P}\hat{\boldsymbol{Y}}_{T+h}, \boldsymbol{S}\boldsymbol{P}\boldsymbol{W}_{T+h}\boldsymbol{P}'\boldsymbol{S}') \tag{3.6.5}$$

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

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

Result 1 can be easily shown as follows. From (3.5.2) the energy score at the upper limit of α is given by, $||\boldsymbol{y}_{T+h} - \boldsymbol{S}\boldsymbol{P}\hat{\boldsymbol{y}}_{T+h}||^2$. Then the expectation of energy score with respect to the true distribution is equivalent to the trace of mean squared forecast error, i.e.

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

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

It should be noted that $\boldsymbol{W}_{T+h}$ can be estimated in different methods which yields different estimation of $\boldsymbol{R}'_{\perp}$. Table 3.2 summarizes these methods.

It is worth mentioning that all these forecasting methods were well established in the context of point forecast reconciliation (Hyndman, Lee, and Wang, 2016; Hyndman et al., 2011; Wickramasuriya, Athanasopoulos, and Hyndman, 2018). However our attempt is to emphasis the use of these reconciliation methods in the context of probabilistic forecasts at least in the Gaussian framework.

Simulation setup

We consider the hierarchy given in figure (1) for this simulation study. This hierarchy consists two aggregation levels with four bottom level series. Each bottom-level series will be generated first and add them up to obtain the data for respective upper-level series. Hierarchical time series in practice contain much noisier series in the bottom level than in aggregate series. In order to simulate this feature in the hierarchy, we refer to Wickramasuriya, Athanasopoulos, and Hyndman (2018) and the data generating process will be given as follows.

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

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

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

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

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

and the total series will be obtained as:

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

To get less noisier aggregate series than disaggregate series, we choose Σ, σ_u^2 and σ_v^2 such that,

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

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

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

This follows,

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

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

As such we generate data for the hierarchy with sample size $T = 501$. Then univariate $ARIMA$ models were fitted for each series independently using the first 500 observations and obtain 1-step ahead base (incoherent) forecasts. We use *forecast* package in **R**-software Hyndman (2017) for model fitting and forecasting. Further, different estimates of $\boldsymbol{W}_{T+1}$ and the corresponding $\boldsymbol{R}_\perp'$ were obtained as summarized in Table 3.2. This process was then replicated using 1000 different data sets from the same data generating process.

To assess the predictive performance of different forecasting methods, we use scoring rules as discussed in Section 3.5. In addition to that we use Skill score (Gneiting and Raftery, 2007) for any comparison. For a given forecasting method, evaluated by a particular scoring rule $S(\cdot)$, the skill score will be calculated as follows,

$$Ss[S_B(\cdot)] = \frac{S_B(\boldsymbol{Y}, \boldsymbol{y})^{\mathrm{ref}} - S_B(\check{\boldsymbol{Y}}, \boldsymbol{y})}{S_B(\boldsymbol{Y}, \boldsymbol{y})^{\mathrm{ref}}} \times 100\%, \tag{3.6.6}$$

Table 3.3: *Comparison of incoherent forecasts using bottom level series. The "Skill score" columns give the percentage skill score with reference to the bottom up forecasting method. A positive entry in these columns shows the percentage increase of score for different reconciliation methods with relative to the bottom up method.*

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

Table 3.4: *Comparison of incoherent vs coherent forecasts for the aggregate series using Skill score. "Incoherent" row represent the average score for incoherent forecasts. Each entry above this row represent the percentage skill score with reference to the incoherent forecasts. A positive(negative) entry shows the percentage increase(decrease) of score for different forecasting methods with relative to incoherent forecasts.*

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

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

As it was mentioned before we wish to establish the importance of reconciliation methods from this simulation study. In particular, we compare different reconciliation methods over the conventional bottom-up method and also evaluate the predictive ability of coherent forecasts over incoherent forecasts. For the former comparison, we use bottom level probabilistic forecasts and calculate the percentage skill score based on energy score, log score and variogram score for each reconciliation method with reference to the bottom up method (presented in Table 3.3). For the latter comparison, we use percentage skill score based on CRPS and univariate log score for coherent probabilistic forecasts of each individual series with reference to incoherent forecasts (presented in Tables 3.4 and 3.5).

It is clearly evident from the results in Table 3.3 that the multivariate reconciled forecasts for the bottom level series from MinT(Shrink) and MinT(Sample) outperform the bottom-up forecasts. Further, these two methods produce probabilistic forecasts with best predictive ability in comparison to incoherent forecasts (from Tables 3.4 and 3.5). Moreover, it turns out that OLS and bottom-up methods produce the worst forecasts.

Table 3.5: *Comparison of incoherent vs coherent forecasts for the individual bottom level series using Skill score.*

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

3.7 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 the 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.

Chapter 4

Probabilistic forecast reconciliation in the non-parametric framework

4.1 Introduction

In some time series applications, it is difficult to assume a parametric distribution for the forecast density. Thus non-parametric approaches have been used in forecasting literature. For example, Nicolau (2010) proposed a method based on time and state domain smoothing. Manzan and Zerom (2008) proposed a bootstrap based non-parametric method to obtain density forecasts of time series. This method involves estimating the density forecasts using bootstrapped in-sample errors. Further Vilar, Alonso, and Vilar (2010) used a bootstrap based approach for multivariate density forecasts in clustering time series.

I propose a bootstrap based non-parametric approach to obtain coherent density forecasts for hierarchical time series. This approach is more applicable for the hierarchies where a parametric approach is not preferable.

4.2 Methodology

This section explains the methodology of the proposed the non-parametric bootstrap approach for obtaining coherent probabilistic forecasts in hierarchical time series.

Suppose we fit univariate time series models for the series at each node in the hierarchy, by using past observations up to time T. Using these fitted models, we can generate h step ahead future sample paths for each node, by conditioning on past observations. In order to capture the dependency structure of the hierarchy, we incorporate, bootstrapped training errors from the fitted models. That is, let $\boldsymbol{\Gamma}_{(T\times n)} = (\boldsymbol{e}_1, \boldsymbol{e}_2, ..., \boldsymbol{e}_T)'$ denote the in-sample residual matrix where, $\boldsymbol{e}_t = \boldsymbol{y}_t - \hat{\boldsymbol{y}}_t$ is a vector that consists of residuals in each node at time t and stacked in the same order as $\boldsymbol{y}_t$. $\boldsymbol{\Gamma}$ is also referred to as the matrix of in-sample errors. Then we block bootstrap a sample of size h from $\boldsymbol{\Gamma}$, and these bootstrapped errors will be incorporated as the error series for simulating future paths. Taking blocked bootstrapped in-sample errors in generating future paths will implicitly model the dependency structure of the hierarchy.

The simulated future sample paths will be then formed in a vector $\boldsymbol{y}^b_{T+h}$ by stacking the sample paths at each node in the same order as $\boldsymbol{y}_t$. Then these sample paths will be reconciled using point forecast reconciliation methods described in chapter 3, to obtain coherent future paths. i.e. we take,

$$\tilde{\boldsymbol{y}}^b_{T+h} = \boldsymbol{S}(\boldsymbol{R}'_{\perp}\boldsymbol{S})^{-1}\boldsymbol{R}'_{\perp}\boldsymbol{y}^b_{T+h}, \tag{4.2.1}$$

where, $\tilde{\boldsymbol{y}}^b_{T+h}$ denote h-step-ahead reconciled future paths, and $\tilde{\boldsymbol{y}}^b_{T+h} \in \mathbb{C}^m$. Thousands of such bootstrapped reconciled future paths will form an empirical coherent forecast distribution of the hierarchy that lies in the coherent subspace $\mathbb{C}^m$. Notice that, different estimates of $\boldsymbol{R}_\perp$, as discussed in table 3.2 will provide alternative estimates of reconciled future paths. Further, we can obtain bottom-up based future paths by simply aggregating the future paths of bottom level series for their respective upper levels. This is referred to as bottom-up future paths. Even though the bottom-up future paths generates coherent probabilistic forecasts, it cannot be considered as a reconciliation method since it contradicts with the definition of reconciliation.

The predictive ability of these different forecasting methods in generating coherent probabilistic forecasts will be evaluated in the following Monte-Carlo simulation study.

4.3 Monte-Carlo Simulation

This section will examine the performance of probabilistic forecasts based on non-parametric bootstrap procedure discussed above. Data sample of size $T = 1510$ was generated by considering the same hierarchical structure and the data generating process as described in Chapter 3. A rolling window of size 500 was used as the training set to fit univariate ARIMA models for each series in the hierarchy. Then, future sample paths of size 5000 for forecasting horizons $h =$ to 5 were generated from fitted models. When generating these sample paths, block bootstrapped training errors were incorporated to implicitly model the covariance structure of the hierarchy. Finally, these future paths were reconciled using different reconciliation methods as described in section 4. The training window was moved from one observation and repeated the process until we get 1000 replicates of probabilistic forecasts for each forecast horizons.

Energy score and variogram score were calculated for the bottom level series of coherent forecasts from different forecasting methods. To evaluate the importance of reconciliation, Skill score with respect to ES and VS were calculated, by taking bottom-up as the reference method. Results are presented in table 4.1. Further, the skill scores with respect to individual CRPS were calculated, to evaluate the predictive performance of coherent probabilistic forecasts over incoherent forecasts. These results are presented in table 4.2.

It is clear from the results that all reconciliation methods produce coherent probabilistic forecasts with better predictive ability than conventional bottom-up method. Further, it can be noticed that coherent forecasts from MinT(Shrink) and MinT(Sample) having improved predictive ability than incoherent forecasts. Moreover, the bottom-up method produces the worst forecast, some of which are even less accurate than the incoherent forecasts.

Table 4.1: *Comparison of Reconciled vs Bottom up density forecasts in bootstrap approach. "Bottom-up" row represent the average ES and VS for bottom-up forecasts. Each entry above this row represent the percentage skill score with reference to the bottom-up forecasts. A positive(negative) entry shows the percentage increase(decrease) of score for different forecasting methods with relative to bottom-up forecasts.*

Forecasting	h=1		h=2		h=3		h=4		h=5	
method	ES	VS	ES	VS	ES	VS	ES	VS	ES	VS
MinT(Shrink)	**6.39**	**5.83**	**6.48**	**5.54**	**5.62**	**5.43**	**3.94**	**3.98**	**3.45**	**3.64**
MinT(Sample)	6.39	5.83	6.35	5.19	5.29	4.89	3.47	3.09	3.04	2.73
MinT(WLS)	3.83	0.89	3.89	1.04	3.35	1.09	2.44	1.33	2.06	1.28
OLS	3.35	0	3.24	0.35	2.69	0.27	2.06	0.66	1.73	0.73
Bottom Up	*6.26*	*2.23*	*7.72*	*2.89*	*9.26*	*3.68*	*10.67*	*4.52*	*12.16*	*5.49*

Table 4.2: *Comparison of Coherent vs Incoherent probabilistic in bootstrap approach. "Incoherent" row represent the average CRPS for incoherent forecasts. Each entry above this row represent the percentage skill score with reference to the incoherent forecasts. A positive(negative) entry shows the percentage increase(decrease) of score for different forecasting methods with relative to incoherent forecasts.*

Forecasting method	Total	Series-A	Series-B	Series-AA	Series-AB	Series-BA	Series-BB
				h=1			
MinT(Shrink)	0.28	7.17	6.57	**10.04**	2.93	**7.35**	**0.41**
MinT(Sample)	**0.29**	**7.10**	6.45	10.04	**3.02**	7.22	0.32
MinT(WLS)	−1.78	6.46	5.80	5.17	−0.03	6.37	−0.68
OLS	−6.30	5.62	5.49	4.50	0.09	5.42	−1.59
Bottom Up	−69.68	−5.25	−4.51	0.00	0.00	0.00	0.00
Incoherent	*1.86*	*2.73*	*2.68*	*2.93*	*2.78*	*2.98*	*2.35*
				h=2			
MinT(Shrink)	**0.40**	**7.60**	7.76	**10.92**	**2.43**	**7.18**	**0.10**
MinT(Sample)	0.39	7.57	**7.77**	10.66	2.39	7.05	−0.13
MinT(WLS)	−0.19	6.96	6.57	6.63	−0.85	6.18	−1.61
OLS	−2.35	6.25	6.22	5.78	−0.69	5.27	−3.09
Bottom Up	−36.34	−3.77	−2.49	0.00	0.00	0.00	0.00
Incoherent	*3.62*	*3.88*	*3.60*	*3.77*	*3.30*	*3.75*	*2.45*
				h=3			
MinT(Shrink)	0.50	7.15	**3.85**	**10.32**	**2.13**	**5.41**	−1.06
MinT(Sample)	0.44	**7.16**	3.77	9.99	2.08	5.16	−1.69
MinT(WLS)	**0.72**	6.45	3.36	6.57	−1.43	5.19	−3.28
OLS	−0.13	5.87	3.50	5.68	−1.20	4.52	−5.18
Bottom Up	−19.85	−2.47	−2.35	0.00	0.00	0.00	0.00
Incoherent	*5.53*	*5.03*	*4.51*	*4.69*	*3.80*	*4.59*	*2.55*
				h=4			
MinT(Shrink)	0.60	6.47	**2.61**	**8.49**	**0.76**	2.67	−1.97
MinT(Sample)	0.54	**6.54**	2.58	8.11	0.47	2.34	−3.19
MinT(WLS)	**1.02**	5.68	2.16	6.00	−2.16	**3.17**	−2.47
OLS	0.62	5.27	2.51	5.19	−1.89	2.98	−4.22
Bottom Up	−13.46	−1.33	−1.14	0.00	0.00	0.00	0.00
Incoherent	*7.18*	*6.27*	*5.30*	*5.67*	*4.23*	*5.29*	*2.61*
				h=5			
MinT(Shrink)	0.63	6.30	**0.67**	**7.56**	**0.35**	1.63	−2.01
MinT(Sample)	0.56	**6.39**	0.58	7.15	0.02	1.29	−3.47
MinT(WLS)	**1.24**	5.45	0.62	5.39	−3.10	**2.41**	−3.36
OLS	1.16	5.07	1.20	4.64	−2.78	2.43	−5.33
Bottom Up	−9.44	−0.52	−1.42	0.00	0.00	0.00	0.00
Incoherent	*8.72*	*7.46*	*5.96*	*6.78*	*4.66*	*5.95*	*2.68*

Chapter 5

Application

5.1 Analysis of Australian tourism data

Forecasting the tourism flow is beneficial for marketing and business decisions in the tourism industry of any country. Total tourism flow of a country is naturally disaggregated along a geographical hierarchy. Previous studies have shown that hierarchical forecasting methods will accurately estimate the point forecasts in application to the domestic tourism flow in Australia (Athanasopoulos, Ahmed, and Hyndman, 2009, Hyndman et al., 2011, Wickramasuriya, Athanasopoulos, and Hyndman, 2018). However, it must also be important to have probabilistic forecasts, in order to give a proper description about forecast uncertainty of future tourism flow, which was not considered in the published literature. Thus my attempt in this section is to obtain probabilistic forecasts of domestic tourism flow in Australia by considering the geographical hierarchical structure. In order to achieve this goal, I use the non-parametric bootstrap approach developed in chapter 4.

I considered the number of ”overnight trips” as a measure of domestic tourism flow in Australia. This can be further disaggregated into ”overnight trips” in 7 states, 27 zones and 75 regions in Australia (Refer to table A.1 for more information about the hierarchical structure). Considering this hierarchical structure will allow the decision makers in the tourism industry, to obtain the probabilistic forecasts of domestic tourism flow in the country along with the forecasts of each state, zone, and region.

Data were obtained from the National Visitor Survey (NVS) which were collected from an annual sample of 120,000 Australian residents aged 15 years or more, through telephone interviews. Data form a monthly series starting from January 1998 to December 2016. Therefore in the data set, I have a total of 110 (total Australia, 7 states, 27 zones and 75 regions) series with 228 observations in each series.

Initially, I fit univariate ARIMA models to all series in the hierarchy using a training set of size 100. Then 5000 of future paths were obtained for the forecast horizons, $h = 1$ to 5 by incorporating block bootstrapped training errors. Finally, these future paths were reconciled as described in section 4, to obtain coherent probabilistic forecasts for domestic tourism flow. MinT(Shrink), MinT(WLS) and OLS methods were used for reconciliation. It should be noted that I ignore MinT(Sample) since the sample size of training data set is less than the dimension of the hierarchy. It was also considered the conventional bottom-up approach to obtain coherent sample paths. I moved the training set forward by one observation and repeated the process until I get 124 replicates of probabilistic forecasts for each forecast horizon.

Table 5.1: *Comparison of Reconciled vs Bottom up density forecasts of domestic overnight trips. "Bottom-up" row represent the average ES and VS for bottom-up forecasts. Each entry above this row represent the percentage skill score with reference to the bottom-up forecasts. A positive(negative) entry shows the percentage increase(decrease) of score for different forecasting methods with relative to bottom-up forecasts.*

Forecasting	h=1		h=2		h=3		h=4		h=5	
method	ES	VS	ES	VS	ES	VS	ES	VS	ES	VS
					States					
MinT(Shrink)	**19.58**	**27.95**	**14.65**	**23.12**	**12.05**	**21.25**	**7.76**	**14.09**	**9.50**	**15.22**
MinT(WLS)	4.50	6.58	4.37	6.31	3.32	6.01	1.89	3.52	3.58	4.67
OLS	7.00	3.15	5.58	5.01	2.65	3.31	1.83	−0.66	3.58	2.01
Bottom Up	*345.63*	*271.05*	*348.68*	*281.63*	*355.98*	*290.47*	*367.41*	*300.25*	*364.81*	*299.75*
					Zones					
MinT(Shrink)	**13.11**	**19.30**	**10.94**	**19.48**	**9.40**	**18.24**	**6.99**	**15.10**	**7.62**	**15.26**
MinT(WLS)	2.32	1.99	2.55	2.99	1.95	2.61	1.05	1.28	1.27	1.97
OLS	1.91	−5.82	1.58	−2.36	0.47	−1.78	0.25	−2.07	1.29	−1.57
Bottom Up	*228.70*	*2366.28*	*229.54*	*2440.32*	*234.09*	*2510.51*	*238.98*	*2556.95*	*237.98*	*2552.26*
					Regions					
MinT(Shrink)	**10.43**	**15.96**	**9.03**	**16.11**	**7.87**	**15.05**	**6.28**	**12.81**	**6.96**	**13.21**
MinT(WLS)	1.54	1.43	1.75	2.16	1.30	1.83	0.68	0.95	0.97	1.46
OLS	1.37	−3.01	1.28	−0.77	0.49	−0.90	0.30	−1.08	1.14	−0.18
Bottom Up	*194.19*	*10956.41*	*193.50*	*11145.87*	*196.26*	*11355.30*	*198.25*	*11461.30*	*197.76*	*11438.31*

In order to evaluate the predictive performance of reconciled forecasts over conventional bottom-up forecasts, I calculated the average energy score and variogram score over 124 replicates for each level in the hierarchy. Then the percentage skill scores for these scoring methods were obtained by taking the bottom-up as reference forecasting method. Results are presented in table 5.1. Further, to compare the predictive performance of coherent vs incoherent forecasts, average CRPS for 124 replicates were calculated for individual series, and for each forecast horizon. The percentage skill scores were then calculated with reference to the incoherent probabilistic forecasts. Results are presented in table 5.2 and 5.3.

Moreover, I evaluated the means of these probabilistic forecasts. First, the means of probabilistic forecasts for each series at each forecast horizon was calculated. Then the mean absolute deviation (MAD) was obtained over 124 replicates. Percentage improvement of these coherent means relative to incoherent means were calculated and presented in table 5.4.

It is clear from the results presented in table 5.1 that the reconciled probabilistic forecasts using MinT(Shrink) approach having improved performance than all other reconciliation methods. Further, it outperforms the conventional bottom-up method with more than 12% performance increment as measured by the variogram score. MinT(WLS) also performing fairly well, especially in the aggregate levels. MinT(Shrink) also outperform incoherent probabilistic forecasts as evaluated by individual CRPS. It is further important to notice that the means of these reconciled forecasts from MinT(Shrink) method outperform the means of the other probabilistic forecasts.

Table 5.2: *Comparison of Coherent vs Incoherent density forecasts of domestic overnight trips in total Australia and States. "Incoherent" row represent the average CRPS for incoherent forecasts. Each entry above this row represent the percentage skill score with reference to the incoherent forecasts. A positive(negative) entry shows the percentage increase(decrease) of score for different forecasting methods with relative to incoherent forecasts.*

Forecasting method	Australia	NSW	Victoria	Queensland	South Australia	Western Australia	Tasmania	Northern Territory
				h=1				
MinT(Shrink)	**8.07**	**9.61**	**16.84**	**22.38**	**23.38**	**11.25**	**13.38**	**20.41**
MinT(WLS)	−12.31	−7.54	2.64	4.95	1.12	1.31	6.93	2.55
OLS	1.39	2.96	5.14	7.68	1.58	−0.48	−13.04	−22.4
Bottom Up	−20.00	−16.22	−1.12	3.07	−4.64	−6.52	4.43	−0.96
Incoherent	*478.81*	*188.48*	*159.49*	*136.27*	*51.29*	*52.97*	*35.23*	*24.27*
				h=5				
MinT(Shrink)	**1.72**	**6.65**	**11.96**	**6.30**	**14.61**	**3.66**	**16.54**	**24.58**
MinT(WLS)	−6.83	−0.30	2.01	2.21	2.26	−1.77	1.40	6.45
OLS	−0.29	2.20	3.41	4.23	9.50	2.58	9.99	−4.52
Bottom Up	−9.07	−3.33	0.24	1.48	0.44	−7.35	−2.59	4.70
Incoherent	*552.45*	*217.09*	*170.55*	*145.01*	*54.99*	*59.59*	*38.76*	*33.87*

Table 5.3: *Comparison of Coherent vs Incoherent density forecasts of domestic overnight trips in Zones and Regions*

Forecasting	Zones					Regions				
method	h=1	h=2	h=3	h=4	h=5	h=1	h=2	h=3	h=4	h=5
MinT(Shrink)	**14.38**	**11.25**	**10.20**	**8.67**	**9.39**	**9.96**	**8.58**	**7.71**	**6.53**	**6.98**
MinT(WLS)	2.88	2.15	1.85	1.76	1.75	1.54	1.73	1.42	0.82	0.94
OLS	1.79	0.55	−0.12	1.00	2.22	−0.11	−0.07	−0.51	−0.20	0.74
Bottom Up	0.21	−0.66	−0.43	0.35	0.26					
Incoherent	*31.75*	*31.87*	*32.71*	*33.73*	*33.49*	*14.36*	*14.45*	*14.70*	*14.87*	*14.81*

Table 5.4: *Comparison of Coherent vs Incoherent mean forecasts of domestic overnight trips. "Incoherent" row represent the average MAD for incoherent forecasts. Each entry above this row represent the percentage improvement of the forecasting method with reference to the incoherent forecasts. A positive(negative) entry shows the percentage increase(decrease) of score for different forecasting methods with relative to incoherent forecasts.*

Forecasting	Australia					States				
method	h=1	h=2	h=3	h=4	h=5	h=1	h=2	h=3	h=4	h=5
MinT(Shrink)	**11.09**	**5.56**	**4.61**	**2.17**	**1.51**	**16.64**	**11.28**	**12.19**	**6.39**	**8.10**
MinT(WLS)	−9.90	−5.33	−2.17	−0.71	−5.48	0.08	0.26	1.31	−0.16	0.41
OLS	1.62	1.53	1.05	1.16	0.12	2.26	0.47	0.43	−0.22	3.06
Bottom Up	−18.88	−12.33	−7.45	−3.66	−8.23	−5.86	−4.64	−2.60	−2.64	−2.16
Incoherent	*669.21*	*697.51*	*744.54*	*812.17*	*764.24*	*128.71*	*132.00*	*138.92*	*142.92*	*142.07*
	Zones					Regions				
	h=1	h=2	h=3	h=4	h=5	h=1	h=2	h=3	h=4	h=5
MinT(Shrink)	**14.64**	**11.15**	**10.71**	**8.58**	**9.16**	**10.87**	**8.81**	**8.05**	**6.71**	**6.88**
MinT(WLS)	2.77	2.13	1.92	1.59	1.68	1.75	1.80	1.54	0.88	0.95
OLS	1.59	0.55	0.04	0.55	1.80	−0.07	−0.17	−0.53	−0.41	0.31
Bottom Up	−0.09	−0.90	−0.66	0.26	0.28					
Incoherent	*43.95*	*44.38*	*45.58*	*47.01*	*46.53*	*20.11*	*20.23*	*20.58*	*20.85*	*20.72*

Chapter 6

Conclusions

While there is a well-established literature on hierarchical point forecast reconciliation, this research proposes doing reconciliation in the probabilistic forecasting framework. Initially, I provided a distinct definition for density forecast reconciliation which allows for both linear and non-linear reconciliation. This was further illustrated for its use in the linear reconciliation of incoherent densities in practice.

Assuming a multivariate Gaussian distribution for the hierarchy, it was shown that, MinT approach is not only optimal in producing mean coherent forecasts but also optimal in producing probabilistic forecasts at least in the Gaussian framework. An extensive Monte Carlo simulation study further showed that the coherent Gaussian forecasts through reconciliation have high predictive ability compared to incoherent Gaussian densities. Moreover, it shows that MinT(Shrink) and MinT(Sample) reconciliation methods outperform conventional bottom-up forecasting method.

I also proposed a novel bootstrap based non-parametric approach to obtain coherent probabilistic forecasts. This method initially involves producing future paths for individual series from the univariate fitted models, by incorporating bootstrapped training errors. Then each of these future paths will be reconciled using point forecast reconciliation methods. Repeating this process to get thousands of such reconciled future paths will give a possible sample from the coherent forecast distribution of the hierarchy. Monte-Carlo simulation study evidence that the reconciled sample paths through MinT(Shrink) and MinT(Sample) have improved predictive ability than incoherent sample paths as well as bottom-up based coherent sample paths.

This non-parametric bootstrap approach was further applied to obtain coherent probabilistic forecasts of domestic tourism flow in Australia. Considering ”overnight trips” as a measure of tourism flow, I compared the predictive ability of probabilistic forecasts from different reconciliation methods. The results clearly evident that the coherent probabilistic forecasts from MinT(Shrink) reconciliation method provide substantially improved forecasts of domestic tourism flow in Australia.

Bibliography

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

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

Dunn, D. M., Williams, W. H., and Dechaine, T. L. (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. M. Poggi, pp. 297–317.

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

Gel, Y., Raftery, A. E., 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, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association* **102**(477), 359–378.

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

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

Gross, C. W. and Sohl, J. E. (1990). Disaggregation methods to expedite product line forecasting. *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, R. J., Lee, A. J., and Wang, E. (2016). Fast computation of reconciled forecasts for hierarchical and grouped time series. *Computational Statistics and Data Analysis* **97**, 16–32.

Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., and Shang, H. L. (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics and Data Analysis* **55**(9), 2579–2589.

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

Kahn, K. B. (1998). *Revisiting top-down versus bottom-up forecasting*. http://search.ebscohost.com/login.aspx?direct=true{\&}db=bth{\&}AN=985713{\&}lang=pt-br{\&}site=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.

Manzan, S. and Zerom, D. (2008). A bootstrap-based non-parametric forecast density. *International Journal of Forecasting* **24**(3), 535–550.

McSharry, P. E., 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.

Nicolau, J. (2010). Nonparametric Density Forecast Based on Time- and State-Domain. **720**(December 2010), 706–720.

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

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

Schäfer, J. and Strimmer, K. (2005). A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics. *Statistical Applications in Genetics and Molecular Biology* **4**(1).

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

Schwarzkopf, A. B., Tersine, R. J., and Morris, J. S. (1988). Top-down versus bottom-up forecasting strategies. *nternational Journal of Production Research* **26**(11), 1833.

Vilar, J. A., Alonso, A. M., and Vilar, J. M. (2010). Non-linear time series clustering based on non-parametric forecast densities. *Computational Statistics and Data Analysis* **54**(11), 2850–2865.

Wickramasuriya, S. L., Athanasopoulos, G., and Hyndman, R. J. (2018). Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *J American Statistical Association*. to appear.

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

Appendix A

Table A.1: *Geographical hierarchy of Australian tourism flow*

	Level 0 - Total	
1	Tot	Australia
	Level 1 - States	
2	A	NSW
3	B	Victoria
4	C	Queensland
5	D	South Australia
6	E	Western Australia
7	F	Tasmania
8	G	Northern Territory
	Level 2 - Zones	
9	AA	Metro NSW
10	AB	North Coast NSW
11	AC	South Coast NSW
12	AD	South NSW
13	AE	North NSW
14	AC	ACT
15	BA	Metro VIC
16	BB	West Coast VIC
17	BC	East Coast VIC
18	BC	North East VIC
19	BD	North West VIC
20	CA	Metro QLD
21	CB	Central Coast QLD
22	CC	North Coast QLD
23	CD	Inland QLD
24	DA	Metro SA
25	DB	South Coast SA
26	DC	Inland SA
27	DD	West Coast SA
28	EA	West Coast WA
29	EB	North WA
30	EC	South WA
31	FA	South TAS
32	FB	North East TAS
33	FC	North West TAS
34	GA	North Coast NT
35	GB	Central NT
	Level 2 - Regions	
36	AAA	Sydney
37	AAB	Central Coast
38	ABA	Hunter
39	ABB	North Coast NSW
40	ABC	Hunter
41	ACA	South Coast
42	ADA	Snowy Mountains
43	ADB	Capital Country
44	ADC	The Murray
45	ADD	Riverina
45	AEA	Central NSW
46	AEB	New England North West
47	AEC	Outback NSW
Zones cont.		
48	AED	Blue Mountains
49	AFA	Canberra
50	BAA	Melbourne
51	BAB	Peninsula
52	BAC	Geelong
53	BBA	Western
54	BCA	Lakes
55	BCB	Grippsland
56	BCD	Phillip Island
57	BDA	Central Murray
58	BDB	Goulburn
59	BDC	High Country
60	BDD	Melbourne East
61	BDE	Upper Yarra
62	BDF	Murray East
63	BEA	Wimmera+Mallee
64	BEB	Western Grampians
65	BEC	Bendigo Loddon
66	BED	Macedon
67	BEE	Spa Country
68	BEF	Ballarat
69	BEG	Central Highlands
70	CAA	Gold Coast
71	CAB	Brisbane
72	CAC	Sunshine Coast
73	CBA	Central Queensland
74	CBB	Bundaberg
75	CBC	Fraser Coast
76	CBD	Mackay
77	CCA	Whitsundays
78	CCB	Northern
79	CCC	Tropical North Queensland
80	CDA	Darling Downs
81	CDB	Outback
82	DAA	Adelaide
83	DAB	Barossa
84	DAC	Adelaide Hills
85	DBA	Limestone Coast
86	DBB	Fleurieu Peninsula
87	DBC	Kangaroo Island
88	DCA	Murraylands
89	DCB	Riverland
90	DCC	Clare Valley
91	DCD	Flinders Range and Outback
92	DDA	Eyre Peninsula
93	DDB	Yorke Peninsula
94	EAA	Australia's Coral Coast
95	EAB	Experience Perth
96	EAC	Australia's South West
97	EBA	Australia's North West
98	ECA	Australia's Golden Outback
Regions cont.		
99	FAA	Hobert and South
100	FBA	East Coast
101	FBB	Launceston, Tamar & North
102	FCA	North West
103	FCB	Wilderness West
104	GAA	Darwin
105	GAB	Kakadu Arnhem
106	GAC	Katherine Daly
107	GBA	Barkly
108	GBB	Lasseter
109	GBC	Alice Springs
110	GBD	MacDonnell