Probabilistic Forecasts for Hierarchical Time Series

Puwasala Gamakumara
Department of Econometrics and Business Statistics,
Monash University,
VIC 3800, Australia.
Email: puwasala.gamakumara@monash.edu
and
Anastasios Panagiotelis*
Department of Econometrics and Business Statistics,
Monash University,
VIC 3800, Australia.
Email: anastasios.panagiotelis@monash.edu
and
George Athanasopoulos
Department of Econometrics and Business Statistics,
Monash University,
VIC 3800, Australia.
Email: george.athanasopoulos@monash.edu
and
Rob J Hyndman
Department of Econometrics and Business Statistics,
Monash University,
VIC 3800, Australia.
Email: rob.hyndman@monash.edu

November 2, 2019

Abstract

TBC

[*] The authors gratefully acknowledge the support of Australian Research Council Grant DP140103220. We also thank Professor Mervyn Silvapulle for valuable comments.

1 Introduction

Large collections of time series often follow some aggregation structure. For example, tourism flows of a country can be disaggregated along a geographic hierarchy of states, zones, and cities. Such collections of time series are generally referred to as hierarchical time series. To ensure aligned decision making, it is important that forecasts across all levels of aggregation add up. This property is called "coherence". If the forecasts are not coherent, then these can be adjusted so that they become coherent. Earlier approaches for obtaining coherent forecasts involve generating first-stage forecasts for series in a single level of the hierarchy and then aggregating these up or disaggregate these down to obtain forecasts for the remaining series. These are often call "bottom-up" and "top-down" forecasts respectively. For example see Dunn et al. (1976), Gross & Sohl (1990) and references therein.

An alternative approach to these single level forecasting methods is to do forecast "reconciliation". Reconciliation starts with a set of incoherent forecasts for the entire hierarchy and then revises these so that they are coherent with the aggregate constraints, see for example Athanasopoulos et al. (2009), Hyndman et al. (2011), Van Erven & Cugliari (2015), Shang & Hyndman (2017). From this literature we see that coherency and reconciliation has been extensively developed for the point forecasting case. Generalising both of these concepts, particularly the latter, to probabilistic forecasting is a gap that we seek to address in this chapter.

In contrast to the point forecasts, the entire probability distribution of future values provides a full description of the uncertainty associated with the predictions (Abramson & Clemen 1995, Gneiting & Katzfuss 2014). Therefore probabilistic forecasting has become of great interest in many disciplines such as, economics (Zarnowitz & Lambros 1987, Rossi

2014), meteorological studies (Pinson et al. 2009, McLean Sloughter et al. 2013), energy forecasting (Wytock & Kolter 2013, Ben Taieb, Huser, Hyndman & Genton 2017) and retail forecasting (Böse et al. 2017). However, the attention on probabilistic forecasts in the hierarchical literature has been limited. Indeed to the best of our knowledge, Ben Taieb, Taylor & Hyndman (2017) and Jeon et al. (2019) are the only papers to deal with probabilistic forecasts in the hierarchical time series. Although Ben Taieb, Taylor & Hyndman (2017) reconcile the means of predictive distributions, the overall distributions are constructed in a bottom-up fashion rather than using a reconciliation approach. Jeon et al. (2019) propose a novel method for probabilistic forecast reconciliation based on cross-validation which is particularly applied to temporal hierarchies. In contrast to these studies, the main objective of this chapter is to generalise both the concepts of coherence and reconciliation from point to probabilistic forecasting.

Extending the geometric interpretation related to point forecast reconciliation derived in (Panagiotelis et al. 2019) we provide new definitions of coherence and forecast reconciliation in the probabilistic setting. We also cover the topic of forecast evaluation of probabilistic forecasts via scoring rules. In particular, we prove that for a coherent data generating process, the log score is not proper with respect to incoherent forecasts. Therefore we recommend the use of the energy score or variogram score for comparing reconciled to unreconciled forecasts. Two or more reconciled forecasts can be compared using log score, energy score or variogram score, although we show that comparisons should be made on the full hierarchy for the latter two scores.

When parametric density assumptions are made we describe how the probabilistic forecast definitions lead to a reconciliation procedure that merely involves a change of basis and marginalisation. We show that probabilistic reconciliation via linear transformations can recover the true predictive distribution as long as the latter is in the elliptical class.

We provide conditions for which this linear transformation is a projection, and although this projection cannot be feasibly estimated in practice, we provide a heuristic argument in favour of MinT reconciliation.

Further we propose a new method to generate coherent forecasts when the parametric distributional assumptions are not applicable. This method uses a non-parametric bootstrap based approach to generate future paths for all series in the hierarchy and then reconcile each sample path using projections. This will provide a possible sample from the reconciled predictive density of the hierarchy. An extensive simulation study was carried out to find the optimal reconciliation of bootstrap future paths with respect to a proper scoring rule. This has shown that the MinT method is at least as good as the optimal method for reconciling future paths.

Finally we applied both parametric and non-parametric approaches to generate probabilistic forecasts for domestic tourism flow in Australia. The results show that reconciliation improves forecast accuracy compared to incoherent forecasts in both parametric and non-parametric approaches and furthermore, MinT reconciliation performs best.

The remainder of the paper is structured as follows. In Section 2.1 notation and some preliminary work on point forecast reconciliation is discussed. Section 2 contains the definitions and interpretation of coherent probabilistic forecasts and reconciliation. In Section 5 we consider the evaluation of probabilistic hierarchical forecasts via scoring rules. Parametric forecast reconciliation and some theoretical results related to elliptical distributions are discussed in Section 3 while the non-parametric approach is introduced in Section 4. An empirical application on tourism forecasting is contained in Section 7. Finally Section 8 concludes with some discussion and thoughts on future research.

2 Hierarchical probabilistic forecasts

The geometric intuition in hierarchical point forecasts provides a solid basis for extending the idea into the probabilistic framework. We start with providing definitions for coherent probabilistic forecasts and probabilistic forecast reconciliation.

2.1 Point Forecasting

Before extending the concepts of coherence and reconciliation to the probabilistic setting we first briefly refresh these concepts in the case of the point forecasts. In do so, we follow the more geometric intepretation introduced by (Panagiotelis et al. 2019).

A *hierarchical time series* is a collection of time series adhering to some known linear constraints. Stacking the value of each series at time t into a vector $\boldsymbol{y}_t$, the constraints imply that $\boldsymbol{y}_t$ lies in an m-dimensional linear subspace of $\mathbb{R}^n$ for all t. This subspace is referred to as the *coherent subspace* and is denoted as $\mathfrak{s}$. A typical (and the original) motivating example is collections of time series some of which are aggregates of other series. In this case $\boldsymbol{b}_t \in \mathbb{R}^m$ can be defined as the values of the *bottom-level series* at time t and the aggregation constraints can be formulated as,

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

where $\boldsymbol{S}$ is an $n \times m$ constant matrix for a given hierarchical structure.

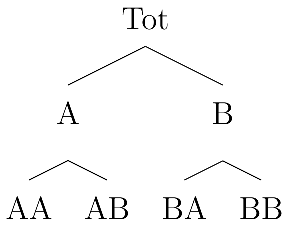

Figure 1: An example of a two level hierarchical structure.

An example of such a hierarchy is depicted in Figure 1. This hierarchy consists of $m = 4$ bottom-level series with $n = 7$ total number of series. Further, $\boldsymbol{y}_t = [y_{Tot,t}, y_{A,t}, y_{B,t}, y_{AA,t}, y_{AB,t}, y_{BA,t}, y_{BB,t}]$ $\boldsymbol{b}_t = [y_{AA,t}, y_{AB,t}, y_{BA,t}, y_{BB,t}]'$ and

$$\boldsymbol{S} = \left(\begin{array}{cccc} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ \multicolumn{4}{c}{\boldsymbol{I}_4} \end{array}\right),$$

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

The connection between this characterisation and the coherent subspace is that the columns of $\boldsymbol{S}$ span $\mathfrak{s}$. Below, it will be convenient at times to think of premultiplication by $\boldsymbol{S}$ as a mapping from $\mathbb{R}^m \rightarrow \mathbb{R}^n$ in which case we use the notation $s(.)$. Finally we note that while $\boldsymbol{S}$ depends on defining coherence in terms of bottom level series, and alternative definitions will lead to a different $\boldsymbol{S}$ matrix, the columns of all such matrices will span the same coherent subspace $\mathfrak{s}$.

When forecasting hierarchical time series there a number of reasons why forecasts may not adhere to constraints. In this case forecasts are called *incoherent* and denoted $\hat{\boldsymbol{y}}_{t+h}$, with the subscript $t + h$ implying a h step made at time t. To exploit the fact that the target of the forecast adheres to known linear constraints, these forecasts can be adjusted in a process known as *forecast reconciliation.* At its most general, this involves selecting a mapping $\psi : \mathbb{R}^n \rightarrow \mathfrak{s}$ and then setting $\tilde{\boldsymbol{y}}_{t+h} = \psi(\hat{\boldsymbol{y}}_{t+h})$, where $\tilde{\boldsymbol{y}}_{t+h} \in \mathfrak{s}$ is called the *reconciled* forecast. This mapping itself may be considered as the composition of two mappings $\psi = s \circ g$. Here, $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ produces a new set of bottom level forecasts using the forecasts of all series, and these are then aggregated via s. When both mappings are linear this corresponds to premultiplying by a matrix $\boldsymbol{SG}$. Several choices of $\boldsymbol{G}$ are currently extant in the literature, including the bottom-up (Dunn et al. 1976), OLS, WLS and MinT (Hyndman et al. 2011, Wickramasuriya et al. 2019) methods. These are special cases where $s \circ g$ is a projection. As an extension, Panagiotelis et al. (2019) also consider a bias correction step, which can be thought of as applying a translation followed by a projection.

2.2 Coherent probabilistic forecasts

We now turn our attention towards defining the concept of coherence in a probabilistic setting. The following method bears some resemblance to defining coherence in the point setting by selecting bottom level series and defining an appropriate $\boldsymbol{S}$ matrix. First let $(\mathbb{R}^m, \mathscr{F}_{\mathbb{R}^m}, \nu)$ be a probability triple, where $\mathscr{F}_{\mathbb{R}^m}$ is the usual Borel σ-algebra on $\mathbb{R}^m$. This triple can be thought of as a probabilistic forecast for the bottom level series. A sigma-algebra $\mathscr{F}_{\mathfrak{s}}$ can then be constructed as the collection of sets $s(\mathcal{B})$ for all $\mathcal{B} \in \mathscr{F}_{\mathbb{R}^m}$, where $s(\mathcal{B})$ denotes the image of $\mathcal{B}$ under the mapping $s(.)$.

Definition 2.1 (Coherent Probabilistic Forecasts)**.** Given the triple, $(\mathbb{R}^m, \mathscr{F}_{\mathbb{R}^m}, \nu)$, a coherent probability triple is given by $\mathfrak{s}$, the sigma algebra $\mathscr{F}_{\mathfrak{s}}$ and a measure $\breve{\nu}$ such that

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

To the best of our knowledge, the only other definition of coherent probabilistic forecasts is given by Ben Taieb, Huser, Hyndman & Genton (2017) who define coherent probabilistic forecasts in terms of convolutions. According to their definition, probabilistic forecasts are coherent when a convolution of forecast distributions of disaggregate series is identical to the forecast distribution of the corresponding aggregate series. While these definitions do not contradict one another we believe our definition has two advantages. First it can more naturally be easily be extended to problems with non-linear constraints with the coherent subspace $\mathfrak{s}$ replaced with a manifold. Second, the geometric understanding of coherence facilitates a definition of probabilistic forecast reconciliation to which we now turn our attention.

2.3 Probabilistic forecast reconciliation

Let $(\mathbb{R}^n, \mathscr{F}_{\mathbb{R}^n}, \hat{\nu})$ be a probability triple characterising a probabilistic forecast for all n series. The hat is used for $\hat{\nu}$ analogously with $\hat{\boldsymbol{y}}$ in the point forecasting case. The objective is to derive a $\hat{\nu}$, a measure that assigns probability to each element of the σ-algebra $\mathscr{F}_{\mathfrak{s}}$.

Definition 2.2. The reconciled probability measure of $\hat{\nu}$ with respect to the mapping $\psi(.)$ is a probability measure $\tilde{\nu}$ on $\mathfrak{s}$ with σ-algebra $\mathscr{F}_{\mathfrak{s}}$ such that

$$\tilde{\nu}(\mathcal{A}) = \hat{\nu}(\psi^{-1}(\mathcal{A})) \qquad \forall \mathcal{A} \in \mathscr{F}_{\mathfrak{s}}\,,$$

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

This definition is a natural extension of the notion of forecast reconciliation to the probabilistic setting. In the point forecasting case, the reconciled point forecast is obtained by passing an incoherent point forecast through a transformation. Similarly, if the incoherent probabilistic forecast assigns a probability to a set of points, then the reconciled probabilisitic forecast will assign the same probability to the same points after they have been passed through a transformation. The transformation ψ can also be expressed as a composition of two transformations $s \circ g$. In this case, an m-dimensional reconciled probabilistic distribution ν can be obtained such that $\nu(\mathcal{B}) = \hat{\nu}(g^{-1}(\mathcal{B}))$ for all $\mathcal{B} \in \mathscr{F}_{\mathbb{R}^m}$ and a probabilistic forecast for the full hierarchy can then be obtained via Definition 2.1

This definition of probabilistic forecast reconciliation can be applied to any mapping continuous mapping ψ, where continuity is required to ensure that the open sets in $\mathbb{R}^n$ used to construct $\mathscr{F}_{\mathbb{R}^n}$ are mapped to open sets in $\mathfrak{s}$. However, from this point on, we restrict our attention to the case where ψ is a linear mapping. This is depicted in Figure 2 for the case when ψ is a projection. This figure is only a 2-dimensional schematic, since even the most trivial hierarchy is 3-dimensional. In this figure, the arrow labelled $\boldsymbol{S}$ spans an m-dimensional coherent subspace $\mathfrak{s}$, while the arrow labelled $\boldsymbol{R}$ spans an $n-m$-dimensional direction of projection. The mapping g collapses all points in the blue shaded region $g^{-1}(\mathcal{B})$ to the black interval $\mathcal{B}$ which has an image under s given by the red interval $s(\mathcal{B})$. Under our definitions of coherence and reconciliation the same probability is assigned to the red shaded region under the reconciled measure as is assigned to the blue region under the incoherent measure.

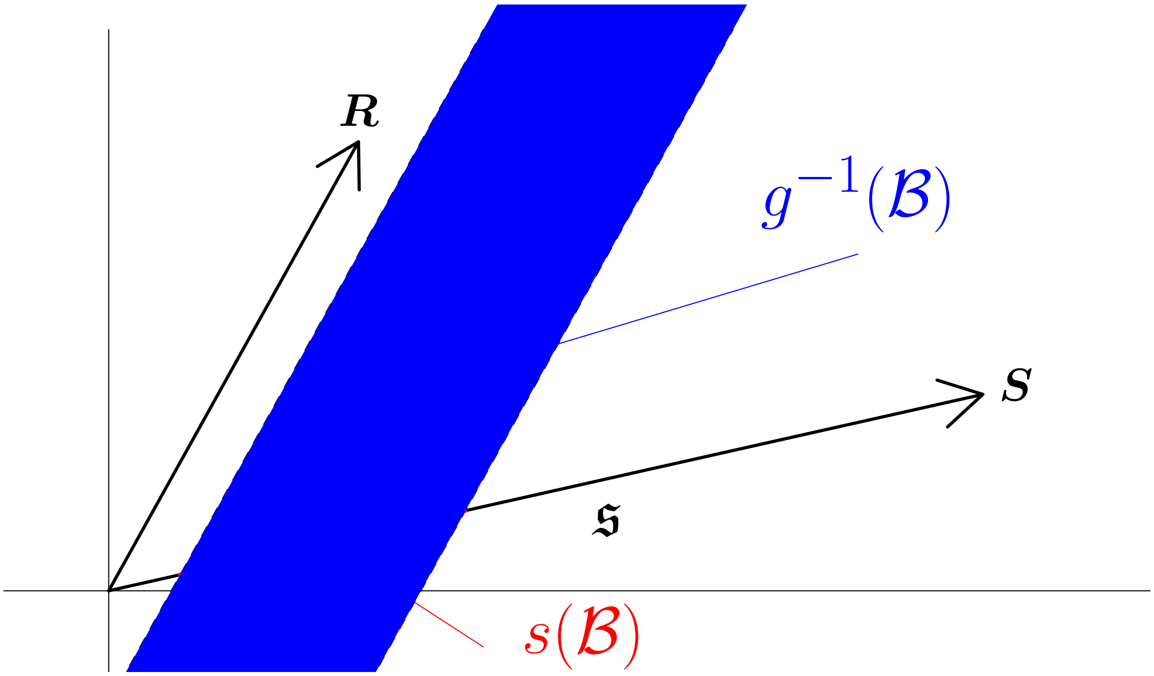

Figure 2: Summary of probabilistic forecast reconciliation. The probability that $\boldsymbol{y}_{t+h}$ lies in the red line segment under the reconciled probabilistic forecast is defined to be equal to the probability that $\boldsymbol{y}_{t+h}$ lies in the shaded blue area under the unreconciled probabilistic forecast. Note that since the smallest possible hierarchy involves three dimensions, this figure is only a schematic.

3 Analytical solution

In this section we describe how a reconciled distribution can be derived analytically, from an incoherent (or base) multivariate probabilistic distribution. We restrict our attention to the case where s and g are linear mappings, in which case reconciliation involves a change of coordinates and marginalisation. This is particularly useful when the base probabilistic forecast is in the elliptical class since the elliptical class is closed under a change of coordinates and marginalisation meaning the reconciled distribution can be obtained analytically.

For ease of exposition we begin by considering the probability that the bottom level series $\boldsymbol{b}$ lies in a region $\mathcal{I}$ a region that is the Cartesian product of intervals $(l_1, u_1), (l_2, u_2), \dots (l_m, u_m)$. We denote this as $\boldsymbol{l} \succ \boldsymbol{b} \succ \boldsymbol{u}$ where $\boldsymbol{l} = (l_1, l_2, \dots, l_m)$, $\boldsymbol{u} = (u_1, u_2, \dots, u_m)$ and $\succ$ denotes element-wise inequality between vectors. The pre-image of $\mathcal{I}$ under g can similarly be denoted as all points $\boldsymbol{y}$ satisfying $\boldsymbol{l} \succ \boldsymbol{G}\boldsymbol{y} \succ \boldsymbol{u}$. By Definition 2.2

$$\Pr(\boldsymbol{b} \in \mathcal{I}) = \int\limits_{\boldsymbol{l} \succ \boldsymbol{G}\boldsymbol{y} \succ \boldsymbol{u}} \hat{f}(\boldsymbol{y}) d\boldsymbol{y}\,,$$

where $\hat{f}$ is the density of the base probabilistic forecast. Now consider a change of variables to an n-dimensional vector $\boldsymbol{z}$ where $y = \boldsymbol{\Gamma}\boldsymbol{z}$. Here, $\boldsymbol{\Gamma} = \left(\boldsymbol{G}^{-} \vdots \boldsymbol{G}_{\perp}\right)$ where $\boldsymbol{G}^{-}$ is an $n \times m$ pseudo inverse of $\boldsymbol{G}$ such that $\boldsymbol{G}\boldsymbol{G}^{-} = \boldsymbol{I}$ and $\boldsymbol{G}_{\perp}$ is an $n \times (n-m)$ orthogonal complement of $\boldsymbol{G}$ such that $\boldsymbol{G}\boldsymbol{G}_{\perp} = \boldsymbol{0}$. By the change of variables

$$
\begin{aligned}
\Pr(\boldsymbol{b} \in \mathcal{I}) &= \int\limits_{\boldsymbol{l} \succ \boldsymbol{G}\boldsymbol{y} \succ \boldsymbol{u}} \hat{f}(\boldsymbol{y}) d\boldsymbol{y} \\
&= \int\limits_{\boldsymbol{l} \succ \boldsymbol{G}\boldsymbol{\Gamma}\boldsymbol{z} \succ \boldsymbol{u}} \hat{f}(\boldsymbol{\Gamma}\boldsymbol{z})|\boldsymbol{\Gamma}| d\boldsymbol{z} \\
&= \int\limits_{\boldsymbol{l} \succ \boldsymbol{z}_1 \succ \boldsymbol{u}} \hat{f}(\boldsymbol{\Gamma}\boldsymbol{z})|\boldsymbol{\Gamma}| d\boldsymbol{z}\,,
\end{aligned}
$$

where $\boldsymbol{b}$ denotes the first m elements of $\boldsymbol{z}$. Also letting $\boldsymbol{a}$ denote the last n-m elements of $\boldsymbol{z}$ the integral above can be written as

$$
\Pr(\boldsymbol{b} \in \mathcal{I}) = \int\limits_{\boldsymbol{l} \succ \boldsymbol{b} \succ \boldsymbol{u}} \int \hat{f}(\boldsymbol{G}^{-}\boldsymbol{b} + \boldsymbol{G}_{\perp}\boldsymbol{a})|\boldsymbol{\Gamma}| d\boldsymbol{a} d\boldsymbol{b}
$$

From the above, the density of the reconciled probabilistic forecast for the bottom level series is given by

$$
\tilde{f}(\boldsymbol{b}) = \int \hat{f}(\boldsymbol{G}^{-}\boldsymbol{b} + \boldsymbol{G}_{\perp}\boldsymbol{a})|\boldsymbol{\Gamma}| d\boldsymbol{a} \tag{2}
$$

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

$$
f(\hat{\boldsymbol{y}}) = f(\boldsymbol{S}\tilde{\boldsymbol{b}} + \boldsymbol{R}\tilde{\boldsymbol{a}})|(\boldsymbol{S}\ \boldsymbol{R})| \tag{3}
$$

The expression $\hat{\nu}(g^{-1}(\mathcal{B}))$ in Definition 2.2 is equivalent to the probability statement $\Pr(\hat{\boldsymbol{y}} \in$

$g^{-1}(\mathcal{B}))$. After the change of basis, this is equivalent to $\Pr(\tilde{\boldsymbol{b}} \in \mathcal{B})$, which implies

$$\begin{aligned}\Pr(\hat{\boldsymbol{y}} \in g^{-1}(\mathcal{B})) &= \int\limits_{g^{-1}(\mathcal{B})} f(\hat{\boldsymbol{y}})d\hat{\boldsymbol{y}} &\quad (4)\\ &= \int\limits_{\mathcal{B}}\int f(\boldsymbol{S}\tilde{\boldsymbol{b}} + \boldsymbol{R}\tilde{\boldsymbol{a}})|(\boldsymbol{S}\ \boldsymbol{R})|d\tilde{\boldsymbol{a}}d\tilde{\boldsymbol{b}}. &\quad (5)\end{aligned}$$

After integrating out over $\tilde{\boldsymbol{a}}$, a step analogous to setting $\tilde{\boldsymbol{a}} = 0$ for point forecasting, we obtain an expression that gives the probability that the reconciled bottom-level series lies in the region $\mathcal{B}$. This corresponds to $\nu(\mathcal{B})$ in Definition 2.2. To make a valid probability statement about the entire hierarchy we simply use the bottom-level probabilistic forecasts together with Definition 2.1.

Example: Gaussian Distributions

Suppose an unreconciled probabilistic forecast is Gaussian with mean $\hat{\boldsymbol{\mu}}$ and variance-covariance matrix $\hat{\boldsymbol{\Sigma}}$. Let the unreconciled density be given by

$$f(\hat{\boldsymbol{y}}) = (2\pi)^{-n/2}|\hat{\boldsymbol{\Sigma}}|^{-1/2}\exp\left\{-\frac{1}{2}\left[(\hat{\boldsymbol{y}} - \hat{\boldsymbol{\mu}})'\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{y}} - \hat{\boldsymbol{\mu}})\right]\right\}. \quad (6)$$

In an alternative basis,

$$f(\tilde{\boldsymbol{b}}, \tilde{\boldsymbol{a}}) = (2\pi)^{-\frac{n}{2}}\left|\hat{\boldsymbol{\Sigma}}\right|^{-\frac{1}{2}}\left|(\boldsymbol{S}\ \boldsymbol{R})\right|\exp\{-\frac{1}{2}q\}, \quad (7)$$

where

$$q = (\boldsymbol{S}\tilde{\boldsymbol{b}} + \boldsymbol{R}\tilde{\boldsymbol{a}} - \hat{\boldsymbol{\mu}})'\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{S}\tilde{\boldsymbol{b}} + \boldsymbol{R}\tilde{\boldsymbol{a}} - \hat{\boldsymbol{\mu}}). \quad (8)$$

The quadratic form q can be rearranged as

$$\begin{aligned}q &= \left((\boldsymbol{S}\ \boldsymbol{R})\begin{pmatrix}\tilde{\boldsymbol{b}}\\ \tilde{\boldsymbol{a}}\end{pmatrix} - \hat{\boldsymbol{\mu}}\right)'\hat{\boldsymbol{\Sigma}}^{-1}\left((\boldsymbol{S}\ \boldsymbol{R})\begin{pmatrix}\tilde{\boldsymbol{b}}\\ \tilde{\boldsymbol{a}}\end{pmatrix} - \hat{\boldsymbol{\mu}}\right),\\ &= \left(\begin{pmatrix}\tilde{\boldsymbol{b}}\\ \tilde{\boldsymbol{a}}\end{pmatrix} - (\boldsymbol{S}\ \boldsymbol{R})^{-1}\hat{\boldsymbol{\mu}}\right)'\left[(\boldsymbol{S}\boldsymbol{R})^{-1}\hat{\boldsymbol{\Sigma}}\left((\boldsymbol{S}\ \boldsymbol{R})^{-1}\right)'\right]^{-1}\left(\begin{pmatrix}\tilde{\boldsymbol{b}}\\ \tilde{\boldsymbol{a}}\end{pmatrix} - (\boldsymbol{S}\ \boldsymbol{R})^{-1}\hat{\boldsymbol{\mu}}\right).\end{aligned}$$

Recall that

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

Then q can be rearranged further as

$$\begin{aligned} q &= \left[\begin{pmatrix} \tilde{\boldsymbol{b}} \\ \tilde{\boldsymbol{a}} \end{pmatrix} - \begin{pmatrix} \boldsymbol{G} \\ \boldsymbol{H} \end{pmatrix} \hat{\boldsymbol{\mu}}\right]' \left[\begin{pmatrix} \boldsymbol{G} \\ \boldsymbol{H} \end{pmatrix} \hat{\boldsymbol{\Sigma}} \begin{pmatrix} \boldsymbol{G} \\ \boldsymbol{H} \end{pmatrix}'\right]^{-1} \left[\begin{pmatrix} \tilde{\boldsymbol{b}} \\ \tilde{\boldsymbol{a}} \end{pmatrix} - \begin{pmatrix} \boldsymbol{G} \\ \boldsymbol{H} \end{pmatrix} \hat{\boldsymbol{\mu}}\right] \\ &= \begin{pmatrix} \tilde{\boldsymbol{b}} - \boldsymbol{G}\hat{\boldsymbol{\mu}} \\ \tilde{\boldsymbol{a}} - \boldsymbol{H}\hat{\boldsymbol{\mu}} \end{pmatrix}' \left[\begin{pmatrix} \boldsymbol{G} \\ \boldsymbol{H} \end{pmatrix} \hat{\boldsymbol{\Sigma}} \begin{pmatrix} \boldsymbol{G} \\ \boldsymbol{H} \end{pmatrix}'\right]^{-1} \begin{pmatrix} \tilde{\boldsymbol{b}} - \boldsymbol{G}\hat{\boldsymbol{\mu}} \\ \tilde{\boldsymbol{a}} - \boldsymbol{H}\hat{\boldsymbol{\mu}} \end{pmatrix} . \end{aligned}$$

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

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

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

$$f(\tilde{\boldsymbol{b}}) = (2\pi)^{-\frac{m}{2}} \left|\boldsymbol{G}\hat{\boldsymbol{\Sigma}}\boldsymbol{G}'\right|^{\frac{1}{2}} \exp\Big\{ -\frac{1}{2}(\tilde{\boldsymbol{b}} - \boldsymbol{G}\hat{\boldsymbol{\mu}})'(\boldsymbol{G}\hat{\boldsymbol{\Sigma}}\boldsymbol{G}')^{-1}(\tilde{\boldsymbol{b}} - \boldsymbol{G}\hat{\boldsymbol{\mu}})\Big\}. \tag{9}$$

This implies that the reconciled probabilistic forecast for the bottom-level series is $\tilde{\boldsymbol{b}} \sim \mathcal{N}(\boldsymbol{G}\hat{\boldsymbol{\mu}}, \boldsymbol{G}\hat{\boldsymbol{\Sigma}}\boldsymbol{G}')$. The reconciled probabilistic forecasts for the whole hierarchy follow a degenerate Gaussian distribution with mean $\boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{\mu}}$ and rank deficient covariance matrix $\boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{\Sigma}}\boldsymbol{G}'\boldsymbol{S}'$.

3.1 Elliptical distributions

We now show that the true predictive distribution can be recovered for elliptical distributions by linear reconciliation via pre-multiplication and translation respectively by a matrix we denote $\boldsymbol{G}_{opt}$ and vector we denote $\boldsymbol{d}_{opt}$. Here, for any square matrix $\boldsymbol{C}$, $\boldsymbol{C}^{1/2}$ and $\boldsymbol{C}^{-1/2}$ are defined to satisfy $\boldsymbol{C}^{1/2}(\boldsymbol{C}^{1/2})' = \boldsymbol{C}$ and $\boldsymbol{C}^{-1/2}(\boldsymbol{C}^{-1/2})' = \boldsymbol{C}^{-1}$, for example $\boldsymbol{C}^{1/2}$ may be obtained via the Cholesky or eigenvalue decompositions.

Theorem 3.1 (Reconciliation for Elliptical Distributions)**.** *Let an unreconciled probabilistic forecast come from the elliptical class with location parameter $\hat{\boldsymbol{\mu}}$ and scale matrix $\hat{\boldsymbol{\Sigma}}$. Let the true predictive distribution of $\boldsymbol{y}$ also belong to the elliptical class with location parameter $\boldsymbol{\mu}$ and scale matrix $\boldsymbol{\Sigma}$. Then the linear reconciliation mapping $g(\breve{\boldsymbol{y}}) = \boldsymbol{G}_{opt}\breve{\boldsymbol{y}} + \boldsymbol{d}_{opt}$ with $\boldsymbol{G}_{opt} = \boldsymbol{a}\hat{\boldsymbol{\Sigma}}^{-1/2}$ and $\boldsymbol{d}_{opt} = \boldsymbol{\mu} - \boldsymbol{S}\boldsymbol{G}_{opt}\hat{\boldsymbol{\mu}}$ recovers the true predictive density where $\boldsymbol{a}$ is any $m \times n$ matrix such that $\boldsymbol{a}\boldsymbol{a}' = \boldsymbol{\Omega}$ and $\boldsymbol{\Omega}$ is a sub-matrix of $\boldsymbol{\Sigma}$ corresponding to the bottom-level series.*

Proof. Since elliptical distributions are closed under affine transformations, and are closed under marginalisation, reconciliation of an elliptical distribution yields an elliptical distribution (although the unreconciled and reconciled distributions may be different members of the class of elliptical distributions). The scale matrix of the reconciled forecast is given by $\boldsymbol{S}\boldsymbol{G}_{opt}\hat{\boldsymbol{\Sigma}}\boldsymbol{G}'_{opt}\boldsymbol{S}'$, while the location matrix is given by $\boldsymbol{S}\boldsymbol{G}_{opt}\hat{\boldsymbol{\mu}} + \boldsymbol{d}_{opt}$. The reconciled scale matrix is

$$\tilde{\boldsymbol{\Sigma}}_{opt} = \boldsymbol{S}\boldsymbol{a}\hat{\boldsymbol{\Sigma}}^{-1/2}\hat{\boldsymbol{\Sigma}}\left(\hat{\boldsymbol{\Sigma}}^{-1/2}\right)'\boldsymbol{a}'\boldsymbol{S}' = \boldsymbol{S}\boldsymbol{\Omega}\boldsymbol{S}' = \boldsymbol{\Sigma}.$$

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

$$\tilde{\boldsymbol{\mu}}_{opt} = \boldsymbol{S}\boldsymbol{G}_{opt}\hat{\boldsymbol{\mu}} + \boldsymbol{\mu} - \boldsymbol{S}\boldsymbol{G}_{opt}\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}.$$

□

A number of insights can be drawn from this theorem. First, although a linear function $g(.)$ can be used to recover the true predictive in the elliptical case, the same does not hold in general. Second, $g(.)$ is not, in general, a projection matrix. The conditions for which a the true predictive density can be recovered by a projection are given below.

Theorem 3.2 (True predictive via projection)**.** *Assume that the true predictive distribution is elliptical with location $\boldsymbol{\mu}$ and scale $\boldsymbol{\Sigma}$. Consider reconciliation via a projection $g(\boldsymbol{y}) = (\boldsymbol{R}'_{\perp}\boldsymbol{S})^{-1}\boldsymbol{R}'_{\perp}\boldsymbol{y}$. The true predictive distribution can be recovered via reconciliation of an elliptical distribution with location $\hat{\boldsymbol{\mu}}$ and scale $\hat{\boldsymbol{\Sigma}}$ when the following conditions hold:*

$$sp(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \subset sp(\boldsymbol{R}) \tag{10}$$

$$sp(\hat{\boldsymbol{\Sigma}}^{1/2} - \boldsymbol{\Sigma}^{1/2}) \subset sp(\boldsymbol{R}) \tag{11}$$

$$\tag{12}$$

Proof. The reconciled location vector will be given by

$$\begin{aligned}\tilde{\boldsymbol{\mu}} &= \boldsymbol{S}(\boldsymbol{R}'_{\perp}\boldsymbol{S})^{-1}\boldsymbol{R}'_{\perp}\hat{\boldsymbol{\mu}}\\ &= \boldsymbol{S}(\boldsymbol{R}'_{\perp}\boldsymbol{S})^{-1}\boldsymbol{R}'_{\perp}\left(\hat{\boldsymbol{\mu}} + \boldsymbol{\mu} - \boldsymbol{\mu}\right)\\ &= \boldsymbol{S}(\boldsymbol{R}'_{\perp}\boldsymbol{S})^{-1}\boldsymbol{R}'_{\perp}\boldsymbol{\mu} + \boldsymbol{S}(\boldsymbol{R}'_{\perp}\boldsymbol{S})^{-1}\boldsymbol{R}'_{\perp}\left(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\right).\end{aligned}$$

Since $\boldsymbol{S}(\boldsymbol{R}'_{\perp}\boldsymbol{S})^{-1}\boldsymbol{R}'_{\perp}$ is a projection onto $\mathfrak{s}$ and $\boldsymbol{\mu} \in \mathfrak{s}$, the first term simplifies to $\boldsymbol{\mu}$. If $\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}$ lies in the span of $\boldsymbol{R}$, then multiplication by $\boldsymbol{R}'_{\perp}$ reduces the second term to $\boldsymbol{0}$. By a similar argument it can be shown that $\tilde{\boldsymbol{\Sigma}}^{1/2} = \boldsymbol{\Sigma}^{1/2}$. The closure property of elliptical distributions under affine transformations ensures that the full true predictive distribution can be recovered. □

Although these conditions will rarely hold in practice and only apply to a limited class of distributions, they do provide some insight into selecting a projection for reconciliation. If the value of $\hat{\boldsymbol{\mu}}$ were equi-probable in all directions, then a projection orthogonal to $\mathfrak{s}$ would be a sensible choice for $\boldsymbol{R}$ since it would in some sense represent a 'median' direction for $\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}$. However, the one-step-ahead in-sample errors are usually correlated suggesting that $\hat{\boldsymbol{\mu}}$ is more likely to fall in some directions than others. Therefore an orthogonal projection after transformation by the inverse of the one-step-ahead in-sample error covariance matrix may be more intuitively appealing. This is exactly what the MinT projection provides, and we demonstrate this in a simulation setting in the following subsection.

4 Sample based solution: A novel non-parametric bootstrap approach

Often in practice we come across hierarchical time series that have high level of disaggregation and/or contain even discrete data. For these time series, parametric distributional assumptions are misleading. An alternative for such cases is to apply non-parametric approaches. Hence we propose a novel non-parametric bootstrap based approach for obtaining coherent probabilistic forecasts.

Our proposed method initially involves obtaining probabilistic forecasts without considering the aggregation constraints. These incoherent probabilistic forecasts are then reconciled to make them coherent. We first focus on the methodology for obtaining base forecasts.

4.1 Incoherent probabilistic forecasts

First we fit appropriate univariate models for each series in the hierarchy based on the training data $\boldsymbol{y}_{1:T}$. We then compute 1-step-ahead training errors as $e_{i,t} = y_{i,t} - \hat{y}_{i,t}$ for

$i = 1, \dots, n$ and $t = 1, \dots, T$ where $\hat{y}_{i,t} = E(y_{i,t}|y_{i,1:t-1})$. The training errors are stored in a matrix $\mathbf{\Gamma}_{(T\times n)} = (\boldsymbol{e}_1, \dots, \boldsymbol{e}_T)'$ where $\boldsymbol{e}_t = (e_{1,t}, \dots, e_{n,t})$ is stored in the same order as $\boldsymbol{y}_t$ for $t = 1, \dots, T$. Next we block bootstrap a sample of size H from $\mathbf{\Gamma}_{\boldsymbol{(T\times n)}}$. That is, we randomly select H consecutive rows from $\mathbf{\Gamma}$ and store in a matrix $\mathbf{\Gamma}^b_{(H\times n)} = (\boldsymbol{e}^b_1, \dots, \boldsymbol{e}^b_H)'$ and repeat this for $b = 1, \dots, B$.

Finally we generate the h-step-ahead future paths using the fitted univariate models conditioning on the past observations. We also incorporate the bootstrapped training errors as the error series for generating these future paths. By doing so we implicitly model the contemporaneous correlation structure of the hierarchy. Further the use of consecutive (block) training errors will ensure that the serial correlation of the series is accounted for. To explain this process more explicitly consider the following example.

Example: Suppose we fit an $ARMA(p,q)$ model for the i^{th} series of the hierarchy. i.e.,

$$\begin{aligned}
y_{i,t} &= \alpha_1 y_{i,t-1} + \alpha_2 y_{i,t-2} + \cdots + \alpha_p y_{i,t-p} + \beta_1\epsilon_{i,t-1} + \beta_1\epsilon_{i,t-2} + \cdots + \beta_q\epsilon_{i,t-q} + \epsilon_{i,t},\\
y_{i,t} &= (\alpha_1 + \alpha_2 L + \cdots + \alpha_p L^{p-1})y_{i,t-1} + (\beta_1 + \beta_1 L + \cdots + \beta_q L^{q-1})\epsilon_{i,t-1} + \epsilon_{i,t}
\end{aligned}$$

where L is the usual lag operator. Then the h-step-ahead b^{th} future path conditional on past information up to and including time t, for the i^{th} series is produced as,

$$\hat{y}^b_{i,t+h} = (\hat{\alpha}_1 + \hat{\alpha}_2 L + \cdots + \hat{\alpha}_p L^{p-1})y_{i,t+h-1} + (\hat{\beta}_1 + \hat{\beta}_1 L + \cdots + \hat{\beta}_q L^{q-1})\epsilon_{i,t+h-1} + e^b_{i,h}$$

where, $e^b_{i,h}$ is the $(h \times i)^{\text{th}}$ element from $\mathbf{\Gamma}^b$,

$$y_{i,t+h-1} = \begin{cases} y_{i,1} : y_{i,T} & \text{for } t+h-1 \leq T \\ \hat{y}^b_{i,T+1} : \hat{y}^b_{i,T+h-1} & \text{for } t+h-1 > T \end{cases}$$

and

$$\epsilon_{i,t+h-1} = \begin{cases} \epsilon_{i,1} : \epsilon_{i,T} & \text{for } t+h-1 \leq T \\ e^b_{i,1} : e^b_{i,h-1} & \text{for } t+h-1 > T \end{cases}.$$

Once we obtain the h-step-ahead sample path for all n series in the hierarchy, we stack them in the same order as $\hat{\boldsymbol{y}}_{t+h}$. Repeating the same process for $b = 1, \dots, B$ we obtain a set of h-step-ahead bootstrapped future paths of size B. We denote this as $\hat{\boldsymbol{\Upsilon}}_{T+h} = (\hat{\boldsymbol{y}}^1_{T+h}, \dots, \hat{\boldsymbol{y}}^B_{T+h})'$ where the b^{th} row of $\hat{\boldsymbol{\Upsilon}}_{T+h}$ represents the h-step-ahead b^{th} sample path for all series in the hierarchy.

We note that $\hat{\boldsymbol{\Upsilon}}_{T+h}$ is an empirical sample from the incoherent probability distribution of the hierarchy. Since the aggregation constraints are not imposed while generating $\hat{\boldsymbol{\Upsilon}}_{T+h}$, it is very unlikely that they lie on the coherent subspace. Thus it requires reconciliation to which we now turn our attention.

4.2 Reconciliation of incoherent future paths

To reconcile the incoherent sample paths, we follow the definition of reconciliation. We project each sample path in $\hat{\boldsymbol{\Upsilon}}_{T+h}$ to the coherent subspace via the projection $\boldsymbol{SG}$. i.e. for any $\boldsymbol{G}$ we can write,

$$\tilde{\boldsymbol{y}}^b_{T+h} = \boldsymbol{SG}\hat{\boldsymbol{y}}^b_{T+h}, \tag{13}$$

consequently we have,

$$\tilde{\boldsymbol{\Upsilon}}'_{T+h} = \boldsymbol{SG}\hat{\boldsymbol{\Upsilon}}'_{T+h}, \tag{14}$$

where, each row in $\tilde{\boldsymbol{\Upsilon}}_{T+h}$ represent a single reconciled sample path. Further $\tilde{\boldsymbol{\Upsilon}}_{T+h}$ form an empirical sample from the reconciled forecast distribution of the hierarchy. Any $\boldsymbol{G}$ matrix introduced in point forecast reconciliation (also given in Table **??**) can be used for this sample path reconciliation. However, in the following subsection we discuss a method to find $\boldsymbol{G}$ that is optimal for probabilistic forecasts with respect to a proper scoring rule.

4.3 Optimal reconciliation of incoherent future paths

Let us now propose to find an optimal $\boldsymbol{G}$ for reconciling future paths by minimising a proper multivariate scoring rule. The respective objective function can be written as,

$$\underset{\boldsymbol{G}_h}{\operatorname{argmin}} \quad \mathrm{E}_{\boldsymbol{Q}}[S(\boldsymbol{S}\boldsymbol{G}_h \hat{\boldsymbol{\Upsilon}}'_{T+h}, \boldsymbol{y}_{T+h})], \tag{15}$$

where S is a proper scoring rule that follows equation (18). We use the subscript h on $\boldsymbol{G}$ to emphasise distinct $\boldsymbol{G}$ matrices for different forecast horizons. Recall that the energy score given in equation (21) is a proper scoring rule. Let $\alpha = 1$ and following equation (22) we can write,

$$\mathrm{ES}(\boldsymbol{S}\boldsymbol{G}_h \hat{\boldsymbol{\Upsilon}}'_{T+h}, \boldsymbol{y}_{T+h}) \approx \frac{1}{B}\sum_{b=1}^{B} ||\boldsymbol{S}\boldsymbol{G}_h \hat{\boldsymbol{y}}^b_{T+h,j} - \boldsymbol{y}_{T+h}|| - \frac{1}{2(B-1)}\sum_{b=1}^{B-1} ||\boldsymbol{S}\boldsymbol{G}_h(\hat{\boldsymbol{y}}^b_{T+h,j} - \hat{\boldsymbol{y}}^{b+1}_{T+h,j})||. \tag{16}$$

where B is the empirical sample size from the coherent forecast distribution. Now we can rewrite the objective function in (15) as,

$$\underset{\boldsymbol{G}}{\operatorname{argmin}} \frac{1}{N}\sum_{j=1}^{N} \left\{ \frac{1}{B}\sum_{b=1}^{B} ||\boldsymbol{S}\boldsymbol{G}_h \boldsymbol{y}^b_{T+h,j} - \boldsymbol{y}_{T+h,j}|| - \frac{1}{2(B-1)}\sum_{b=1}^{B-1} ||\boldsymbol{S}\boldsymbol{G}_h(\boldsymbol{y}^b_{T+h,j} - \boldsymbol{y}^{b+1}_{T+h,j})|| \right\} \tag{17}$$

where, the expectation $E_{\boldsymbol{Q}}$ over true forecast distribution $\boldsymbol{Q}$ is approximated through the sample mean over $\{\mathrm{ES}(\boldsymbol{S}\boldsymbol{G}_h \hat{\boldsymbol{\Upsilon}}'_{T+h,1}, \boldsymbol{y}_{T+h,1}), \ldots, \mathrm{ES}(\boldsymbol{S}\boldsymbol{G}_h \hat{\boldsymbol{\Upsilon}}'_{T+h,N}, \boldsymbol{y}_{T+h,N})\}$. We can use numerical optimization methods to estimate the matrix $\boldsymbol{G}_h$ that minimises the above objective function and thus obtain the optimally reconciled future paths.

We have also considered different parameterisation methods for estimation optimal $\boldsymbol{G}_h$. This was discussed in Section A in Appendix.

5 Evaluation of hierarchical probabilistic forecasts

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

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

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

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

$$\mathrm{E}_{\boldsymbol{Q}}[S(Q, \omega)] \leq \mathrm{E}_{\boldsymbol{Q}}[S(P, \omega)], \tag{18}$$

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

In the context of probabilistic forecast reconciliation there could be two motivations for using scoring rules. The first is to compare unreconciled densities to reconciled densities. Reconciliation itself is a valuable goal since it can be important in aligning decision making across, for example, different units of an enterprise. In the point forecasting literature, forecast reconciliation has also been shown to improve forecast performance (Athanasopoulos et al. 2017, Wickramasuriya et al. 2019). It will be worthwhile to see whether the same holds in the probabilistic forecasting case. The second motivation for using scoring rules is to compare two or more sets of reconciled probabilistic forecasts to one another. The objective here is to evaluate which reconciliation mapping $g(.)$ works best in practice.

5.1 Univariate scoring rules

One way to evaluate hierarchical probabilistic forecasts is via the application of univariate scoring rules to each time series in the hierarchy. A summary can be taken of the expected scores across each margin, for example a mean or median. We consider two such scoring rules. The log score is given by the log density, in this case for each margin of the probabilistic forecast. The continuous rank probability score generalises mean square error and is given by

$$\begin{aligned}\mathrm{CRPS}(\breve{F}_i, y_i) &= \int \Big(\breve{F}_i(\breve{Y}_i) - \mathbb{1}(\breve{Y}_i < y_i)\Big)\, d\breve{Y}_i \qquad (19)\\ &= \mathrm{E}_{\breve{Y}_i}|\breve{Y}_i - y_i| - \frac{1}{2}\mathrm{E}_{\breve{Y}_i}|\breve{Y}_i - \breve{Y}_i^*|\,, \qquad (20)\end{aligned}$$

where $\breve{F}_i$ is the cumulative distribution function of the i^{th} margin of the probabilistic forecast, $\breve{Y}_i$ and $\breve{Y}_i^*$ are independent copies of a random variable with distribution $\breve{F}_i$, and

y_i is the outcome of the i^{th} margin. The expectations in the second line can be approximated by Monte Carlo when a sample from the predictive distribution is available.

An advantage of this approach is that it allows the forecaster to separately evaluate different levels and individual series of the hierarchy to determine where the gains from reconciliation are greatest. For this reason this approach has been used in the limited literature on probabilistic forecasting for hierarchies (Ben Taieb, Huser, Hyndman & Genton 2017, Jeon et al. 2019) to date. A major shortcoming of this approach however, is that evaluating univariate scores on the margins does not account for the dependence in the hierarchy.

5.2 Multivariate scoring rules

While a number of alternative proper scoring rules are available for univariate forecasts, the multivariate case is somewhat more limited. Here we focus on three scoring rules: the log score (LS), the energy score (ES) and the variogram score (VS).

The log score can be approximated using a sample of values from the probabilistic forecast density (Jordan et al. 2017); however it is more commonly used when a parametric form for the density is available for the probabilistic forecast.

The energy score on the other hand can be defined in terms of the characteristic function of the probabilistic forecast, but the following representation in terms of expectations

$$\mathrm{ES}(\breve{\boldsymbol{Y}}_{T+h}, \boldsymbol{y}_{T+h}) = \mathrm{E}_{\breve{\boldsymbol{Y}}}||\breve{\boldsymbol{Y}}_{T+h} - \boldsymbol{y}_{T+h}||^{\alpha} - \frac{1}{2}\mathrm{E}_{\breve{\boldsymbol{Y}}}||\breve{\boldsymbol{Y}}_{T+h} - \breve{\boldsymbol{Y}}^{*}_{T+h}||^{\alpha},\ \alpha \in (0,2]\,, \tag{21}$$

lends itself to easy computation when samples from the probabilistic forecast are available and given as,

$$\mathrm{ES}(\breve{\boldsymbol{Y}}_{T+h}, \boldsymbol{y}_{T+h}) \approx \frac{1}{M}\sum_{i=1}^{M}||\boldsymbol{S}\boldsymbol{G}\breve{\boldsymbol{y}}_{T+h,i} - \boldsymbol{y}_{T+h}|| - \frac{1}{2(M-1)}\sum_{i=1}^{M-1}||\boldsymbol{S}\boldsymbol{G}(\breve{\boldsymbol{y}}_{T+h,i} - \breve{\boldsymbol{y}}_{T+h,i+1})||, \tag{22}$$

where, $\breve{\boldsymbol{y}}_{T+h,i}$ is the i^{th} Monte-Carlo sample from the forecast distribution. An interesting limiting case is where $\alpha = 2$, where it can be easily shown that energy score simplifies to mean squared error around the mean of the predictive distribution. In this limiting case, the energy score is proper but not strictly proper. Pinson & Tastu (2013) also argue that the energy score has low discriminative ability for incorrectly specified covariances, even though it discriminates the misspecified means well.

In contrast, Scheuerer & Hamill (2015) have shown that the variogram score has a higher discrimination ability for misspecified means, variances and correlation structures than the energy score. When $\breve{\boldsymbol{y}}$ is a random variable from probabilistic forecast $\breve{F}$, the empirical variogram score is defined as

$$\mathrm{VS}(\breve{F}, \boldsymbol{y}) = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \left(|y_i - y_j|^p - E_{\breve{Y}_i, \breve{Y}_j} |\breve{Y}_i - \breve{Y}_j|^p \right)^2. \tag{23}$$

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

5.2.1 Comparing unreconciled forecasts to reconciled forecasts

For both reconciled and unreconciled densities it is possible to obtain a density from the probability measures defined in Section 2. Therefore it may seem sensible to compare unreconciled densities to reconciled densities on the basis of log score. However, the following theorem shows that using the log score may fail in the case of multivariate distributions with a degeneracy.

Theorem 5.1 (Impropriety of log score)**.** *When the true data generating process is coherent, then the log score is improper with respect to the class of incoherent measures.*

Proof. Consider a rotated version of hierarchical time series, $\boldsymbol{z}_t = \boldsymbol{U}\boldsymbol{y}_t$, so that the first m elements of $\boldsymbol{z}_t$ denoted $\boldsymbol{z}_t^{(1)}$ are unconstrained, while the remaining $n-m$ elements denoted

$\boldsymbol{z}_t^{(2)}$ equal 0 when the aggregation constraints hold. An example of the $n \times n$ $\boldsymbol{U}$ is the matrix of left singular vectors of $\boldsymbol{S}$.

Consider the case where the true predictive density is $f_1(\boldsymbol{z}_t^{(1)})\mathbb{1}\left(\boldsymbol{z}_t^{(2)} = \boldsymbol{0}\right)$, and we evaluate an incoherent density given by $f_1(\boldsymbol{z}_t^{(1)})f_2(\boldsymbol{z}_t^{(2)})$, where f_2 is highly concentrated around 0 but still non-degenerate. For example, f_2 may be Gaussian with variance $\sigma^2\boldsymbol{I}$ with $\sigma^2 < (2\pi)^{-1}$. The log score under the true data generating process is

$$LS\left(f, \boldsymbol{z}_t^{(1)}\right) = -\log f_1\left(\boldsymbol{z}_t^{(1)}\right),$$

while that of the unreconciled density is

$$\begin{aligned} LS\left(\hat{f}, \boldsymbol{z}_t^{(1)}\right) &= -\log f_1(\boldsymbol{z}_t^{(1)}) - \log f_2(\boldsymbol{z}_t^{(1)}) && (24)\\ &= -\log f_1(\boldsymbol{z}_t^{(1)}) + \frac{n-m}{2}\log(2\pi\sigma^2) && (25)\\ &< -\log f_1(\boldsymbol{z}_t^{(1)}) = LS\left(f, \boldsymbol{z}_t^{(1)}\right). && (26) \end{aligned}$$

After taking expectations $E\left[LS(f, f)\right] > E\left[LS(\hat{f}, f)\right]$, violating the condition in Equation (18) for a proper scoring rule. □

A similar issue also arises when discrete random variables are modelled as if they were continuous, an issue discussed in Section 4.1 of Gneiting & Raftery (2007). This implies that the log score should be avoided when comparing reconciled and unreconciled probabilistic forecasts.

5.2.2 Comparing reconciled forecasts to one another

Coherent probabilistic forecasts can be completely characterised in terms of basis series; if a probabilistic forecast is available for the basis series, then a probabilistic forecast can be recovered for the entire hierarchy via Definition 2.1. This may suggest that it is adequate to merely compare two coherent forecasts to one another using the basis series only. We now show how this depends on the specific scoring rule used.

For the log score, suppose the coherent probabilistic forecast has density $f(\boldsymbol{b})$. The density for the full hierarchy is given by $f(\boldsymbol{y}) = f(\boldsymbol{S}\boldsymbol{b}) = f(\boldsymbol{b})J^{-1}$, where $J = \prod_{j=1}^{m} \lambda_j$ is a pseudo-determinant of the non-square matrix $\boldsymbol{S}$ and λ_j are the non-zero singular values of $\boldsymbol{S}$. Therefore for any coherent density, the log score of the full hierarchy differs from the log score for the bottom-level series by the term $log(J)$. This term depends only on the structure of the hierarchy and is fixed across different reconciliation methods. Therefore if one method achieves a lower expected log score compared to an alternative method using the bottom-level series only, the same ordering is preserved when an assessment is made on the basis of the full hierarchy.

The same property does not hold for all scores in general. For example, the energy score can be expressed in terms of expectations of norms. In general, since norms are invariant under orthogonal rotations, the energy score is also invariant under orthogonal transformations (Székely & Rizzo 2013, Gneiting & Raftery 2007). In the context of two coherent forecasts, the same is true of a semi-orthogonal transformation from a lower dimensional basis series to the full hierarchy. However, when $\boldsymbol{S}$ is the usual summing matrix, it is not semi-orthogonal. Therefore the energy score computed on the bottom-level series will differ from the energy score computed using the full hierarchy and the ordering of different reconciliation methods may change depending on the basis series used. In this case we recommend computing the energy score using the full hierarchy. Although the discussion here is related to energy score, the same logic holds for other multivariate scores, for example the variogram score.

The properties of multivariate scoring rules in the context of evaluating reconciled probabilistic forecasts are summarised in Table 1.

Table 1: Summary of properties of scoring rules in the context of reconciled probabilistic forecasts.

	Coherent v Incoherent	Coherent v Coherent
Log Score	Not proper	Ordering preserved if compared using bottom-level only
Energy/	Proper	Full hierarchy should be used
Variogram Score	Proper	Full hierarchy should be used

5.2.3 Skill scores

To facilitate comparisons between different forecasting methods, we use skill scores (Gneiting & Raftery 2007). The skill score for a given forecast distribution P, with reference to a forecast distribution P_{ref}, evaluated by a particular scoring rule $S()$, is calculated as,

$$SS(P, \omega) = \frac{E_Q[S(P_{ref}, \omega)] - E_Q[S(P, \omega)]}{E_Q[S(P_{ref}, \omega)]} \tag{27}$$

The skill score gives the percentage improvement of the preferred forecasting method relative to the reference method. A negative valued skill score indicates that a method is worse than the reference method, whereas any positive value indicates that the method is superior to the reference method.

6 Simulations

We now present the simulation study carried-out to evaluate the performance of probabilistic forecasts in both parametric and non-parametric setting. Let us first discuss the data generating process.

6.1 Data generating process (DGP)

The data generating process we consider is the hierarchy given in Figure 1, comprising two aggregation levels with four bottom-level series. Each bottom-level series will be generated first, and then summed to obtain the data for the upper-level series.

First $\{w_{AA,t}, w_{AB,t}, w_{BA,t}, w_{BB,t}\}$ 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. The parameters for the AR and MA components are randomly and uniformly generated from $[0.3, 0.5]$ and $[0.3, 0.7]$ respectively. The errors driving the ARIMA processes were generated from Gaussian and non-Gaussian distributions separately. This will allow us to demonstrate the distinct impact of true DGP for parametric and non-parametric reconciliation approaches.

Gaussian errors:

Errors were jointly generated from a normal distribution, and denoted by $\{\varepsilon_{AA,t}, \varepsilon_{AB,t}, \varepsilon_{BA,t}, \varepsilon_{BB,t}\} \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ $\forall t$, where,

$$\mathbf{\Sigma} = \begin{pmatrix} 5.0 & 3.1 & 0.6 & 0.4 \\ 3.1 & 4.0 & 0.9 & 1.4 \\ 0.6 & 0.9 & 2.0 & 1.8 \\ 0.4 & 1.4 & 1.8 & 3.0 \end{pmatrix}. \tag{28}$$

Non-Gaussian errors:

Non-Gaussian errors were generated from a Gumbel copula model with beta margins. Using a copula model helps to impose a non-linear dependence structure among the series. A two dimensional Gumbel copula is given by,

$$C_\theta(u_1, u_2) = exp\{-[(-ln(u_1))^\theta + (-ln(u_2))^\theta]^{1/\theta}\}.$$

We generate random variates $\{u_{AA}, u_{AB}\}$ from $C_{\theta=10}(.)$ and $\{u_{BA}, u_{BB}\}$ from $C_{\theta=8}(.)$ for series $\{AA, AB\}$ and $\{BA, BB\}$ respectively. Next we generate the errors, $\{\varepsilon_{AA}, \varepsilon_{AB}, \varepsilon_{BA}, \varepsilon_{BB}\}$ as the quantiles from beta distributions with shape parameters $\alpha = 1$ and $\beta = 3$ correspond to $\{u_{AA}, u_{AB}, u_{BA}, u_{BB}\}$.

Signal-to-noise ratio:

In practice, hierarchical time series are likely to have relatively noisier series at lower levels of aggregation. We replicate this feature in our simulated data by adding some noise to $\{w_{AA,t}, w_{AB,t}, w_{BA,t}, w_{BB,t}\}$ as described in Section B.1 in Appendix. We denote these nosier bottom-level series by $\{y_{AA,t}, y_{AB,t}, y_{BA,t}, y_{BB,t}\}$.

6.2 Simulation set up for parametric solution

To compare different reconciliation methods in parametric densities we assume a Gaussian predictive distribution for the hierarchy. We choose the Gaussian case due to its analytical tractability which allows for evaluation using all scoring rules (including the log score).

We generate 2000 observations for each series from the Gaussian and non-Gaussian DGP. We ignore the first 500 observations from each series to avoid the impact from initial values. Using a rolling window of $T = 500$ observations, we fit univariate ARIMA models for each series using the `auto.arima()` function in the `forecast` package (Hyndman et al. 2019) in R (R Core Team 2018). Using the fitted models we generate 1 to 3 steps ahead base (incoherent) Gaussian probabilistic forecasts. We estimate the mean and the variance of this incoherent Gaussian density as the h-steps ahead point forecasts $\hat{\boldsymbol{y}}_{t+h}$ and shrinkage estimator for variance covariance matrix of one-step ahead forecast errors $\hat{\boldsymbol{W}}^{\text{shr}}$ respectively. These were then reconciled using different projections summarised in Table **??**. This process was replicated for 1000 times by rolling the window one step at a time.

To assess the predictive performance of different forecasting methods, we use scoring rules as discussed in Section 5.

6.2.1 Results and discussion

Table 2 summarises the forecasting performance of incoherent, bottom-up, OLS, WLS and two MinT reconciliation methods using log score, energy score and variogram score. The top panel refers to the Gaussian DGP whereas the bottom panel refers to the non-Gaussian DGP. Recall that the log score is improper with respect to incoherent forecasts. Therefore we calculate the skill scores with reference to the bottom-up forecasts instead of incoherent forecasts in all cases and leave blank the cell for log score of the incoherent forecasts. Further, all log scores are evaluated on the basis of bottom-level series only, however these only differ from the log scores for the full hierarchy by a fixed constant. Overall, the MinT methods provide the best performance irrespective of the scoring rule, and all methods that reconcile using information at all levels of the forecast improve upon incoherent forecasts. Bottom-up forecasts perform even worse than incoherent forecasts in some cases. These results hold for both the Gaussian as well as the non-Gaussian DGP.

Tables 3 break down the forecasting performance of the different methods by considering univariate scores on each individual margin. We have only presented the results for forecast horizon $h = 1$ and the results for rest of the forecasts horizons are presented in table B.2 in Appendix. Univariate log score and CRPS are considered, while skill scores are computed with the incoherent forecasts as a reference. When broken down in this fashion, irrespective of DGP, the methods based on MinT perform best for most series and outperform bottom-up forecasts in almost all cases.

Table 2: Comparison of coherent forecasts in forecast for $h = 1$ to 3 steps-ahead. All entries shows the percentage skill score with reference to the bottom-up method. The top panel shows results from the Gaussian DGP and bottom panel shows the results from the non-Gaussian DGP. "ES" and "VS" columns give scores based on the joint forecast distribution across the entire hierarchy. The "LS" column gives the log scores of the joint forecast distribution of the bottom-level.

	h=1			h=2			h=3		
Method	ES(%)	VS(%)	LS(%)	ES(%)	VS(%)	LS(%)	ES(%)	VS(%)	LS(%)
				Gaussian DGP					
MinT(Shrink)	**19.48**	**9.78**	**3.16**	**19.57**	**14.16**	**6.53**	**16.47**	**16.56**	**8.34**
MinT(Sample)	**19.48**	9.74	3.09	19.50	14.16	6.51	16.28	16.42	8.09
WLS	18.08	7.21	0.64	17.68	10.97	2.31	14.99	13.17	3.76
OLS	16.01	5.80	-0.79	15.38	8.43	0.05	13.03	10.26	0.82
Bottom up	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Incoherent	11.65	-0.12		10.58	1.71		8.75	3.64	
				Non-Gaussian DGP					
MinT(Shrink)	**15.04**	**0.69**	**4.52**	**16.98**	**1.34**	**4.55**	**18.00**	**0.66**	**4.01**
MinT(Sample)	15.02	0.59	4.40	16.94	1.02	4.30	17.88	0.64	3.42
WLS	12.72	0.00	0.93	14.22	0.41	1.34	15.20	-0.42	0.89
OLS	11.26	0.17	0.65	12.27	0.48	0.47	13.12	-0.24	0.10
Bottom up	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Incoherent	8.47	-2.79		8.94	-2.09		9.20	-3.62	

Table 3: Comparison of incoherent vs coherent forecasts based on the univariate forecast distribution of each series. Each entry represents the percentage skill score with reference to the incoherent forecasts based on “CRPS” and “LS”. These entries show the percentage increase in score for different forecasting methods relative to the incoherent forecasts for $h = 1$ step-ahead forecast. Results from the Gaussian DGP are presented in the top panel whereas the results from the non-Gaussian DGP are presented in the bottom panel

	Total		A		B		AA		AB		BA		BB	
R.method	CRPS	LS	CRPS	LS	CRPS	LS	CRPS	LS	CRPS	LS	CRPS	LS	CRPS	LS
							Gaussian DGP							
MinT(Shrink)	-0.13	-0.01	**9.37**	3.12	5.42	1.67	3.91	1.30	**12.04**	**3.82**	**10.07**	**3.12**	1.47	0.47
MinT(Sample)	-0.08	-0.04	**9.37**	**3.13**	5.24	1.67	**4.12**	**1.38**	11.99	**3.82**	9.90	3.10	**1.57**	**0.51**
WLS	-2.91	-1.24	8.78	2.86	**5.49**	**1.73**	1.10	0.42	10.37	3.21	9.12	2.78	-1.14	-0.26
OLS	-19.22	-6.86	6.28	2.06	4.86	1.58	0.80	0.25	8.47	2.59	7.91	2.39	-1.52	-0.49
Bottom up	-140.27	-33.67	-13.75	-3.89	-11.10	-3.17	0.01	0.00	0.04	0.00	0.15	0.00	-0.09	0.00
Incoherent	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
							Non-Gaussian DGP							
MinT(Shrink)	-1.16	-0.26	**0.92**	0.27	**11.90**	4.91	**3.40**	**1.31**	-0.11	-0.11	**13.22**	**5.00**	**2.37**	0.87
MinT(Sample)	-1.16	-0.72	**0.92**	**0.28**	**11.90**	**4.94**	**3.40**	1.29	-0.11	-0.13	**13.22**	4.99	**2.37**	**0.91**
WLS	**0.01**	**0.35**	-1.02	-0.52	9.95	4.07	2.92	1.20	-1.90	-0.74	8.50	3.13	-0.96	-0.26
OLS	-96.77	-84.90	0.55	0.08	6.48	2.57	2.70	1.07	-1.46	-0.55	6.19	2.25	-0.81	-0.21
Bottom up	-541.40	-246.37	-4.60	-1.87	-8.99	-3.11	-0.10	0.00	-0.12	0.00	-0.02	0.00	-0.08	0.00
Incoherent	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

6.3 Simulation setup for non-parametric solution

We now implement the non-parametric approach introduced in Section 4. We use the same DGP as discussed before. First let us elaborate the simulation set up as follows.

1. Generate time series with 2500 data points for each series in the hierarchy.

2. Consider a rolling window of 600 observations. We call this the "outer" rolling window.

    i. Inside this outer rolling window consider an inner rolling window of $T = 500$ observations.

    ii. For this inner rolling window, fit univariate ARIMA models to each series in the hierarchy.

    iii. Based on these fitted models, generate $B = 1000$ of $h = 1$ to 3 steps-ahead incoherent future paths incorporating bootstrap errors as described in Subsection 4.1. Thus we get $\{\hat{\boldsymbol{\Upsilon}}_{T+1,j=1}, \hat{\boldsymbol{\Upsilon}}_{T+2,j=1}, \hat{\boldsymbol{\Upsilon}}_{T+3,j=1}\}$.

    iv. Repeat step (iii) for $j = 1, \dots, N$ where $N = 100$ by rolling the inner window one step ahead at a time.

    v. Collect $\{\hat{\boldsymbol{\Upsilon}}_{T+h,j=1}, \dots, \hat{\boldsymbol{\Upsilon}}_{T+h,j=100}\}$ for $h = 1, \dots, 3$ into separate arrays of matrices.

    vi. For each forecast horizon h, estimate the optimal $\boldsymbol{G}_h$ that will reconcile $\{\hat{\boldsymbol{\Upsilon}}_{T+h,j=1}, \dots, \hat{\boldsymbol{\Upsilon}}_{T+h,j=100}\}$ by minimising the average energy score as explained in Subsection 4.3. Denote this as $\boldsymbol{G}_h^{Opt}$.

    vii. Roll the inner rolling window another one step ahead and repeat steps (ii) and (iii). Denote these future paths by $\hat{\boldsymbol{\Upsilon}}_{T+h}$ for $h = 1, 2, 3$.

    viii. Compute $\tilde{\boldsymbol{\Upsilon}}'_{T+h} = \boldsymbol{S}\boldsymbol{G}_h\hat{\boldsymbol{\Upsilon}}'_{T+h}$ for $h = 1, 2, 3$ using $\boldsymbol{G}_h^{Opt}$ as well as using other $\boldsymbol{G}$ matrices given in Table **??**.

3. Repeat Step 2 1000 times by rolling the outer rolling window one step-ahead at a time. Collect 1000 reconciled future paths, $\tilde{\boldsymbol{\Upsilon}}_{T+h}$, from different reconciliation methods for $h = 1, 2, 3$ and evaluate the forecasting performances.

6.3.1 Results and discussion

Following this simulation process, we generate reconciled non-parametric probabilistic forecasts for Gaussian as well as non-Gaussian data. To assess their predictive performance we use energy and variogram scores. Results are presented in Table 4.

We use Mann-Whitney test to compare the difference of scores between reconciliation methods. The results support that the ES and VS for all reconciled forecasts are significantly lower than those of incoherent forecasts. This implies that all reconciliation methods produce coherent probabilistic forecasts with improved predictive ability compared to the incoherent forecasts. In addition to that, the MinT(Shrink) and Optimal method have similar prediction accuracy as there is no significant difference between the scores from these reconciliation methods. Results are consistent for both Gaussian and non-Gaussian data.

We have also compute the reconciled probabilistic forecasts from different parameterisations for optimal method. These results are presented in Section B.3 in Appendix. From these results we note that the scores for different optimal methods are equivalent irrespective to the forecast horizon or the DGP, implying that there is no difference in results due to different parameterisations of $\boldsymbol{G}$.

However we note that optimal reconciliation required a high computational cost for larger hierarchies. Further, it requires sufficient data points to learn the $\boldsymbol{G}$ matrix. Thus we suggest using the MinT $\boldsymbol{G}$ for reconciling bootstrapped future paths for two reasons. First, it is computationally efficient relative to the optimal method, and second, it produces accurate probabilistic forecasts that are at least as good as the Optimal method with respect to the energy score.

Table 4: Energy scores (ES) and variogram scores (VS) for probabilistic forecasts from different reconciliation methods are presented. Bottom row represent the scores for base forecasts which are not coherent. The smaller the scores, the better the forecasts are.

	Non-Gaussian DGP						Gaussian DGP					
Reconciliation	h=1		h=2		h=3		h=1		h=2		h=3	
method	ES	VS	ES	VS	ES	VS	ES	VS	ES	VS	ES	VS
Optimal*	5.36	1.21	5.51	1.27	5.83	1.38	9.59	4.86	11.50	5.38	13.80	6.13
MinT(Shrink)*	5.33	1.19	5.50	1.26	5.77	1.34	9.43	4.78	11.40	5.33	13.70	6.09
WLS	5.43	1.23	5.60	1.30	5.89	1.40	9.64	4.93	11.70	5.60	14.10	6.39
OLS	5.51	1.23	5.70	1.30	5.98	1.40	9.91	4.93	12.10	5.60	14.50	6.39
Incoherent	*5.71*	*1.28*	*5.94*	*1.37*	*6.27*	*1.49*	*10.40*	*5.31*	*12.70*	*6.22*	*15.20*	*7.14*

The differences in scores between methods noted by "" are statistically insignificant. The differences between these and the incoherent forecasts are statistically significant.*

7 Application: Forecasting Australian domestic tourism flow

In this section we illustrate how the probabilistic forecast reconciliation methods can be used in practice, by forecasting domestic tourism flows in Australia. Previous studies have shown that reconciliation for this data generate more accurate point forecasts compared to the bottom-up or incoherent forecasts. For example see Athanasopoulos et al. (2009), Hyndman et al. (2011) and Wickramasuriya et al. (2019). This study is the first to apply reconciliation methods for forecasting tourism in a probabilistic framework.

7.1 Data

As a measure of domestic tourism flows, we consider the "overnight trips" to different destinations across the country. Data are collected through the National Visitor Survey (NVS) managed by Tourism Research Australia based on an annual sample of $120,000$ Australian residents aged 15 years or more, through telephone interviews (Tourism Research Australia 2019).

The total number of overnight trips in Australia can be naturally disaggregated through a geographic hierarchy. This hierarchy consists of 7 states [1] in the 1st level of disaggregation, 27 zones in the 2nd level of disaggregation and 76 regions in the bottom-level and thus comprises 110 series in total. More details about the individual series are provided in Table **??**. We consider monthly overnight trips for all series spanning the period January 1998 to December 2018. This gives 152 observations per series.

[1] We have considered ACT as a part of New South Wales and Northern Territory as a state.

7.2 Forecasting methodology

We apply both the parametric and non-parametric reconciliation approaches as discussed in previous sections. We use a rolling window of 100 observations as the training sample where the first training sample will span the period Jan-1998 to Apr-2006. Based on this training set we fit univariate ARIMA and ETS models for each series in the hierarchy using automated functions `auto.arima()` and `ets()` from the `forecast` package (Hyndman et al. 2019) in R software (R Core Team 2018). From the estimated models we generate parametric and non-parametric probabilistic forecasts for one year ahead, i.e for $h = 1, \dots, 12$. For the parametric forecasts, we assume Gaussian densities and obtain the incoherent mean and variance forecasts. These are then reconciled using the methods described in Section 3. For the non-parametric forecasts, we generate the bootstrapped future paths and then reconcile each sample path as described in Section 4. We note that we do not implement the MinT(Sample) approach as the sample size of training data set is less than the dimension of the hierarchy. Using a rolling window, one month at a time, we replicate the process until the end of the sample. This yields, 152 1-step ahead, 151 2-steps ahead through to 141 12-step ahead probabilistic forecasts available for evaluation. We note that we only present the results for ARIMA models in the following section. The results for ETS models are similar and we present these in the Appendix.

7.3 Evaluation, results and discussion

We evaluate the predictive accuracy using scoring rules. More specifically we use energy and variogram scores to assess the predictive accuracy of multivariate forecast distributions across the entire hierarchy as well as for the different disaggregation levels. CRPS is used to assess the predictive accuracy of univariate forecast distributions for each series in the hierarchy. We calculate average scores over the replications for each forecast horizon separately. In the results that follow we present skill scores for each of the coherent predictive distributions with reference to the incoherent distributions. A positive (negative) values in the skill score indicates a gain (loss) in forecast accuracy over the incoherent forecast distribution.

Figure 3 shows the skill scores with respect to the multivariate predictive distributions across the entire hierarchy from the different methods. Figure 4 shows the evaluation across each level. The top panels present the results from the Gaussian approach while the bottom panels present the results from the non-parametric approach. Both figures show that almost all reconciliation methods improve forecast accuracy irrespective of whether the parametric or non-parametric approaches are implemented. Furthermore, the bottom-up approach shows losses compared to the incoherent forecasts at all forecast horizons. This reflects the fact that bottom-level series are noisier and therefore more challenging to forecast. Finally and most importantly, MinT(Shrink) outperforms all probabilistic forecast reconciliation methods for both parametric and non-parametric approaches.

Figure 5 shows the predictive accuracy of the univariate forecast distributions for the Total overnight trips. OLS and MinT(Shrink) reconciliation methods show gains in accuracy for the top level of the hierarchy for both Gaussian and non-parametric approaches.