ISSN 1440-771X

MONASH
BUSINESS
SCHOOL

Department of Econometrics and Business Statistics

http://business.monash.edu/econometrics-and-business-statistics/research/publications

Probabilistic Forecasts in Hierarchical Time Series

Puwasala Gamakumara
Anastasios Panagiotelis
George Athanasopoulos
Rob J Hyndman

July 2018

Working Paper ??/??

Probabilistic Forecasts in Hierarchical Time Series

Puwasala Gamakumara
Department of Econometrics and Business Statistics,
Monash University,
VIC 3800, Australia.
Email: Puwasala.Gamakumara@monash.edu

Anastasios Panagiotelis
Department of Econometrics and Business Statistics,
Monash University,
VIC 3800, Australia.
Email: Anastasios.Panagiotelis@monash.edu

George Athanasopoulos
Department of Econometrics and Business Statistics,
Monash University,
VIC 3800, Australia.
Email: George.Athanasopoulos@monash.edu

Rob J Hyndman
Department of Econometrics and Business Statistics,
Monash University,
VIC 3800, Australia.
Email: Rob.Hyndman@monash.edu

6 July 2018

JEL classification: ??

Probabilistic Forecasts in Hierarchical Time Series

Abstract

Adjusting forecasts to ensure coherence with aggregation constraints that hold for real data - a process known as forecast reconiliation - is extended from point forecasts to probabilistic forecasts. This is achieved by redefining forecast reconciliation in terms of linear mappings in general, and projections more specifically. New theorems establish that the true predictive distribution can be recovered in the elliptical case by a linear mapping and general conditions are derived for when this linear mapping is a projection. A geometric interpretation is also used to prove two new theoretical results for point forecasting; that reconciliation via projection a) preserves unbiasedness and b) dominates unreconciled forecasts in a mean squared error sense. Strategies for forecast evaluation based on scoring rules are discussed and it is shown that the popular log score is an improper scoring rule with respect to the class of unreconciled forecasts when the true predictive disstribution coheres with aggregation constraints. Finally evidence from simulation study show that reconciliation based on an oblique projection, derived from the MinT method of Wickramasuriya, Athanasopoulos, and Hyndman, 2018 for point forecasting, outperforms both reconciled and unreconciled alternatives.

1 Introduction

Large collections of time series often follow some aggregation structure. For example, the electricity demand of a country can be disaggregated according to a geographical hierarchy of states, cities, and individual households or by a temporal hierarchy of yearly, quarterly and monthly demand. In these so-called hierarchical time series, forecasts are often required for all individual series. To ensure aligned decision making it is important for forecasts at the most disaggregated - or bottom-level - to add up to forecasts at more aggregated - or higher - levels. This property is referred to as "coherence". This issue has been extensively covered for point forecasting in a literature we review below, while the case of probabilistic forecasting is a gap that we seek to address in this work.

Traditional approaches to ensure coherent point forecasts produce first-stage forecasts at a single level. To describe these we use the small hierarchy in Figure 1 where the variable labelled Tot is the sum of the series A and series B, the series A is the sum of series AA and series AB and

the series B is the sum of the series BA and BB. In the bottom-up approach (Dunn, Williams, and Dechaine, 1976), forecasts are produced at the most disaggregated level (Series AA, AB, BA and BB) and then summed to recover all top level-series. Alternatively, in the top-down approach (Gross and Sohl, 1990) a top-level forecast is first produced (Series Tot) and bottom level forecasts are recovered as historical or forecasted proportions of the top-level forecast. A middle-out approach is a hybrid between these two that for the hierarchy below would produce first stage forecasts for Series A and B.

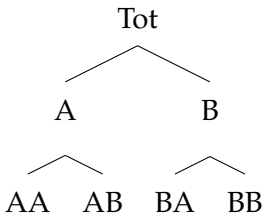

Figure 1: *Two level hierarchical diagram.*

In recent years, reconciliation methods introduced by Hyndman et al. (2011) have become increasingly popular. For these methods, first stage forecasts are independently produced for all series rather than series at a single level. Since these so-called 'base' forecasts are rarely coherent in practice, they are subsequently adjusted or 'reconciled' to ensure coherence. To date reconciliation has typically been formulated as a regression problem with alternative reconciliation methods resembling different least squares estimators. These include Ordinary Least Squares OLS Hyndman et al., 2011, Weighted Least Squares WLS (Athanasopoulos et al., 2017) and a Generalised Least Squares (GLS) estimator (Wickramasuriya, Athanasopoulos, and Hyndman, 2018), more commonly referred to as MinT since it minimises the trace of the squared error matrix. These methods have been shown to outperform traditional alternatives across a range of simulated and real-world datasets (Hyndman et al., 2011; Erven and Cugliari, 2014; Wickramasuriya, Athanasopoulos, and Hyndman, 2018) since they use information at all levels of the hierarchy and in some sense hedge against the risk of model misspecification at a single level.

A shortcoming of the existing literature is a focus on point forecasting despite an increased understanding over the past decade of the importance of providing a full predictive distribution for forecast uncertainty see Gneiting and Katzfuss, 2014, and references therein. Indeed to the best of our knowledge, the as yet unpublished work of Ben Taieb et al., 2017 is the only paper to

deal with the issue of coherent probabilistic forecasts, and although the means of the predictive distributions are reconciled, the overall distributions are constructed in a bottom up fashion. In contrast, the main objective of our paper is to generalise both coherence and reconciliation from point to probabilistic forecasting.

To facilitate the extension of point forecast reconciliation to probabilistic forecasting, we first provide a geometric interpretation of existing point reconciliation methods, framing them in terms of projections. In addition to being highly intuitive, this allows us to establish a number of theoretical results. We prove two new theorems about point forecast reconciliation, the first showing that reconciliation via projections preserves the unbiasedness of base forecasts while the second shows that reconciled forecasts dominate unreconciled forecasts since projections are bounded operators. We provide definitions of coherence and forecast reconciliation in the probabilistic setting and describe how these 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.

We also cover the issue 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. As such we recommend the use of the energy score and 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.

The remainder of the papers is structured as follows. In section 2 coherence is defined geometrically for both point and probabilistic forecasts. Section 3 contains definitions of point and probabilistic forecast reconciliation as well as our main theoretical results. In Section 4 we consider the evaluation of probabilistic hierarchical forecasts via scoring rules while a simulation study comparing unreconciled probabilistic forecasts and different kinds of reconciled probabilistic forecasts is provided in Section 5 and Section 6 concludes.

2 Coherent forecasts

2.1 Notation and Preliminaries

Throughout the paper, we will follow notational conventions used in Wickramasuriya, Athanasopoulos, and Hyndman (2018) as much as possible. A *hierarchical time series* is a collection of n variables where some variables are aggregates of other variables and we let $\boldsymbol{y}_t \in \mathbb{R}^n$ be a vector comprised of observations of all variables in the hierarchy at time t. For example for the hierarchy in Figure 1, $n = 7$ and $\boldsymbol{y}_t = [y_{Tot,t}, y_{A,t}, y_{B,t}, y_{C,t}, y_{AA,t}, y_{AB,t}, y_{BA,t}, y_{BB,t}]'$. The *bottom level series* are defined as those m variables that cannot be formed as aggregates of other variables, we let $\boldsymbol{b}_t \in \mathbb{R}^m$ be a vector comprised of observations of all bottom-level series at time t. For example for the hierarchy in Figure 1, $m = 4$ and $\boldsymbol{b}_t = [y_{AA,t}, y_{AB,t}, y_{BA,t}, y_{BB,t}]'$. The hierarchical structure of the data imply the following holds for all t

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

where $\boldsymbol{S}$ is an $n \times m$ constant matrix that encodes the aggregation constraints. For the hierarchy in Figure 1

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

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

2.2 Coherent Point Forecasts

It is desirable that forecasts, whether point forecasts or probabilistic forecasts, should in some sense respect aggregation constraints. We follow other authors Wickramasuriya, Athanasopoulos, and Hyndman, 2018in using the nomenclature *coherence* to describe this property. FPP?

We now provide new definitions for coherent forecasts in terms of vector spaces that give a geometric understanding of the problem thus facilitating the development of the probabilistic forecast reconciliation in section 3.

Recall that $\boldsymbol{y}_t \in \mathbb{R}^n$ is a n-dimensional time series 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.

Definition 2.1 (Coherent subspace)**.** The m-dimensional linear subspace $\mathfrak{s} \subset \mathbb{R}^n$ that is spanned by the columns of $\boldsymbol{S}$, i.e. $\mathfrak{s} = \text{span}(\boldsymbol{S})$, is defined as the *coherent space*.

Also at times is will be useful to think of pre-multiplication by $\boldsymbol{S}$ as a linear mapping from $\mathbb{R}^m$ to $\mathbb{R}^n$ in which case we use the notation $s(.)$. Although the codomain of $s(.)$ is $\mathbb{R}^n$ its image is the coherent space $\mathfrak{s}$ as depicted in Figure 2.

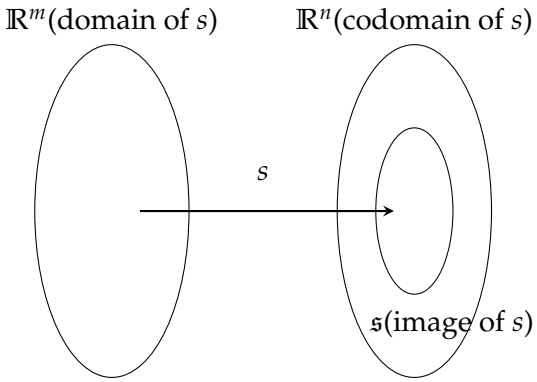

Figure 2: *The domain, codomain and image of the mapping s.*

Definition 2.2 (Coherent Point Forecasts)**.** Let $\breve{\boldsymbol{y}}_{t+h|t} \in \mathbb{R}^n$ be a point forecast of the values of all series in the hierarchy at time $t+h$, made using information up to and including time t. Then $\breve{\boldsymbol{y}}_{t+h|t}$ is *coherent* if $\breve{\boldsymbol{y}}_{t+h|t} \in \mathfrak{s}$.

2.3 Coherent Probabilistic Forecasts

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

Definition 2.3 (Coherent Probabilistic Forecasts)**.** The measure $\breve{\nu}$ is coherent if it has the property

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

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

These definitions of the coherent space $\mathfrak{s}$ and coherent point and probabilistic forecasts are defined in terms of the mapping $s(.)$ and may give the impression that the bottom level series

play an important role in the definition. However, alternative definitions could be formed using any set of basis vectors that spans $\mathfrak{s}$. For example, consider the most simple three variable hierarchy where $y_{1,t} = y_{2,t} + y_{3,t}$. In this case the matrix $\boldsymbol{S}$ has columns $(1,1,0)'$ and $(1,0,1)'$ spanning $\mathfrak{s}$ and premultiplying by $\boldsymbol{S}$ transforms arbitrary values of $y_{2,t}$ and $y_{3,t}$ into a coherent vector for the full hierarchy. However the columns $(1,0,1)'$ and $(0,1,-1)'$ also span $\mathfrak{s}$ and define a mapping that transforms arbitrary values of $y_{1,t}$ and $y_{2,t}$ into a coherent vector for the full hierarchy. The definitions above could be made in terms of any series and not just the bottom level series. In general, we call the series (or linear combinations thereof) used in the definitions of coherence *basis series*. Unless stated otherwise, we will always assume that the basis series are the bottom level series as in Definition 2.2 and Defintion 2.3, since this facilitates comparison with existing approaches in the literature.

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

3 Forecast reconciliation

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

3.1 Point forecast reconciliation

Let $\hat{\boldsymbol{y}}_{t+h|t} \in \mathbb{R}^n$ be any set of incoherent point forecasts at time $t+h$ using information up to and including time t. Let $\boldsymbol{G}$ and $\boldsymbol{d}$ be an $m \times n$ matrix and $m \times 1$ vector respectively and $g : \mathbb{R}^n \to \mathbb{R}^m$ be the mapping $g(\boldsymbol{y}) = \boldsymbol{G}\boldsymbol{y} + \boldsymbol{d}$.

Definition 3.1. The point forecast $\tilde{\boldsymbol{y}}_{t+h|t}$ "reconciles" $\hat{\boldsymbol{y}}_{t+h|t}$ with respect to the mapping $g(.)$ iff

$$\tilde{\boldsymbol{y}}_{t+h|t} = \boldsymbol{S}\left(\boldsymbol{G}\hat{\boldsymbol{y}}_{t+h|t} + \boldsymbol{d}\right) .$$

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

Method	**Estimate of $\boldsymbol{W}_h$**	**Estimate of $\boldsymbol{R}'_\perp$**
OLS	$\boldsymbol{I}$	$\boldsymbol{S}'$
MinT(Sample)	$\hat{\boldsymbol{W}}^{sam}_{T+1}$	$\boldsymbol{S}'(\hat{\boldsymbol{W}}^{sam}_{T+1})^{-1}$
MinT(Shrink)	$\hat{\boldsymbol{W}}^{shr}_{T+1} = \tau\text{Diag}(\hat{\boldsymbol{W}}^{sam}_{T+1}) + (1-\tau)\hat{\boldsymbol{W}}^{sam}_{T+1}$	$\boldsymbol{S}'(\hat{\boldsymbol{W}}^{shr}_{T+1})^{-1}$
WLS	$\hat{\boldsymbol{W}}^{wls}_{T+1} = \text{Diag}(\hat{\boldsymbol{W}}^{shr}_{T+1})$	$\boldsymbol{S}'(\hat{\boldsymbol{W}}^{wls}_{T+1})^{-1}$

Many choices of $g(.)$ currently extant in the literature including the so called OLS Hyndman et al., 2011, WLS and MinTWickramasuriya, Athanasopoulos, and Hyndman, 2018 methods are special cases where $s \circ g$ is a projection and are summarised in Table 1. These can be defined so that $\boldsymbol{G} = (\boldsymbol{R}'_\perp \boldsymbol{S})^{-1}\boldsymbol{S}'$ and $\boldsymbol{d} = \boldsymbol{0}$. Here, $\boldsymbol{R}_\perp$ is a $n \times m$ orthogonal complement to the an $n \times (n-m)$ matrix $\boldsymbol{R}$ where the columns of the latter span the null space of $\mathfrak{s}_\perp$. For example, a straightforward choice of $\boldsymbol{R}$ for the most simple three variable hierarchy where $y_{1,t} = y_{2,t} + y_{3,t}$, is the vector $(1, -1, -1)$ which is orthogonal (in the Euclidean sense) to the columns of $\boldsymbol{S}$. In this case, the matrix $\boldsymbol{R}$ can be interpreted as a 'restrictions' matrix since it has the property that $\boldsymbol{R}'\boldsymbol{y} = \boldsymbol{0}$ for coherent $\boldsymbol{y}$. For the example provided, $\boldsymbol{R}'_\perp = \boldsymbol{S}$ and reconciliation corresponds to the so called 'OLS' method Hyndman et al., 2011. For the case where $\boldsymbol{R}'_\perp \neq \boldsymbol{S}$, for example WLS and MinT, there are two possible interpretations. One is that these are oblique projections in Euclidean space where the columns of $\boldsymbol{R}$ are 'directions' along which incoherent point forecasts are projected onto the coherent space $\mathfrak{s}$. Alternatively, since $\boldsymbol{R}'_\perp$ is usually written in the form $\boldsymbol{S}'\boldsymbol{W}^{-1}$, these projections can be thought of as orthogonal projections after applying the transformation $\boldsymbol{W}^{-1/2}$. A schematic providing a geometric interpretation of point reconciliation is given in Figure 3.

First WLS reference?

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

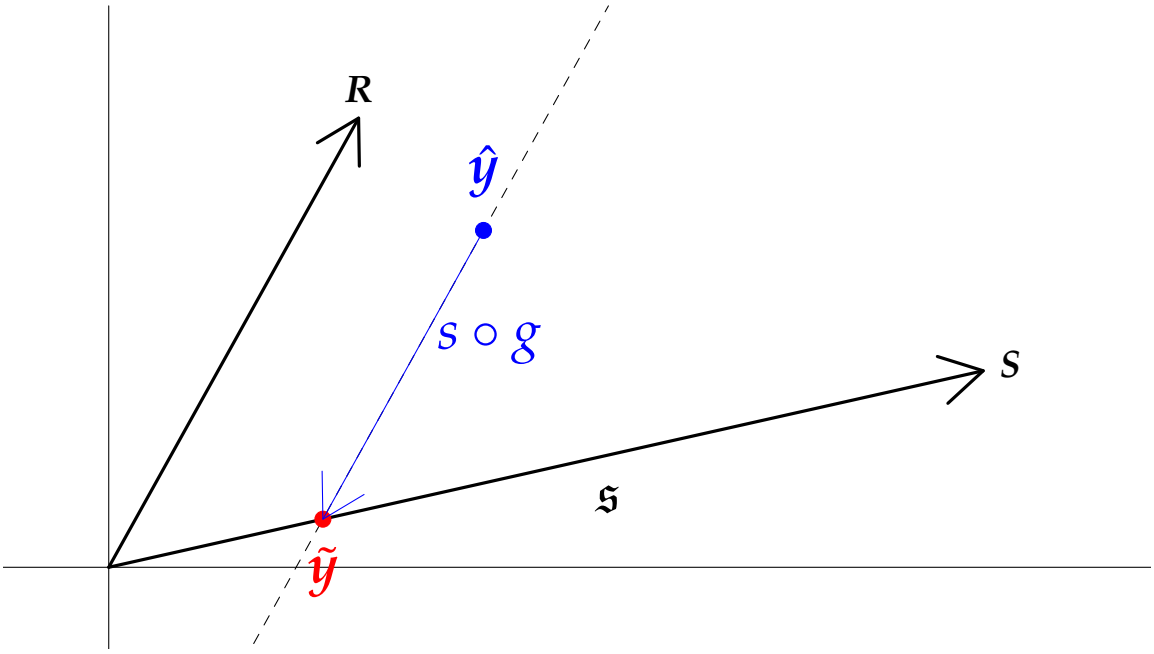

Figure 3: *Summary of probabilistic point reconciliation. The mapping $G(.)$ projects the unreconciled forecast $\hat{\boldsymbol{y}}$ onto Note that since the smallest hierarchy involves three dimensions, this figure is only a schematic*

properties of linear algebra it follows that

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

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

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

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

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

where $S_\perp$ is the orthogonal complements of S. Thus it follows that $\tilde{b}_{t+h} = (R'_\perp S)^{-1} R'_\perp \hat{y}_{t+h}$ and $\tilde{y}_{t+h|t} = S(R'_\perp S)^{-1} R'_\perp \hat{y}_{t+h|t}$. Here $(R'_\perp S)^{-1} R'_\perp$ corresponds to G as defined previously.

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

mention reducing distance to truth?

Before generalising the concept of point reconciliation to probabilistic forecasts we state two theorems that motivate the use of projections for point forecast reconciliation. First, let $\mu_{t+h|t} := E(y_{t+h} | y_1, \ldots, y_t)$ and assume $\hat{y}_{t+h|t}$ is an unbiased prediction, that is $E_{1:t}\left(\hat{y}_{t+h|t}\right) = \mu_{t+h|t}$ where the subscript $1:t$ denotes an expectation taken over the training sample.

Theorem 3.1 (Unbiasedness preserving property)**.** *The reconciled point forecast will also be an unbiased prediction as long as $s \circ g$ is a projection*

Proof. The expected value of the reconciled forecast is given by

$$\begin{aligned} E_{1:t}\left(\tilde{y}_{t+h|t}\right) &= E_{1:t}\left(SG\hat{y}_{t+h|t}\right) \\ &= SGE_{1:t}\left(\hat{y}_{t+h|t}\right) \\ &= SG\mu_{t+h|t} \end{aligned}$$

Since the aggregation constraints hold for the true data generating process $\mu_{t+h|t}$ must lie in $\mathfrak{s}$. If SG is a projection then it is equivalent to the identity map for all vectors that lie in its range. Therefore $SG\mu_{t+h|t} = \mu_{t+h|t}$ when SG is a projection matrix. □

We note the same result does not hold for general G even though the range of $s \circ g$ is $\mathfrak{s}$. Now let y_{t+h} be the realisation of the data generating process at time $t+h$ and let $||v||_2$ be the L2 norm

of vector v. The following theorem shows that reconciliation never increases, and in most cases reduces the sum of squared errors of point forecasts.

Theorem 3.2 (Distance reducing property)**.** *If $\tilde{\boldsymbol{y}}_{t+h|t} = \boldsymbol{SG}\hat{\boldsymbol{y}}_{t+h|t}$ where $\boldsymbol{G}$ is such that $\boldsymbol{SG}$ is an orthogonal projection then the following inequality holds*

$$||(\tilde{\boldsymbol{y}}_{t+h|t} - \boldsymbol{y}_{t+h})||_2^2 \leq ||(\hat{\boldsymbol{y}}_{t+h|t} - \boldsymbol{y}_{t+h})||_2^2$$

Proof. Since the aggregation constraints must hold for all realisations, $\boldsymbol{y}_{t+h} \in \mathfrak{s}$ and $\boldsymbol{y}_{t+h} = \boldsymbol{SG}\boldsymbol{y}_{t+h}$ whenever $\boldsymbol{SG}$ is a projection. Therefore

$$\begin{aligned} ||(\tilde{\boldsymbol{y}}_{t+h|t} - \boldsymbol{y}_{t+h})||_2 &= ||(\boldsymbol{SG}\hat{\boldsymbol{y}}_{t+h|t} - \boldsymbol{SG}\boldsymbol{y}_{t+h})||_2 \\ &= ||\boldsymbol{SG}(\hat{\boldsymbol{y}}_{t+h|t} - \boldsymbol{y}_{t+h})||_2 \end{aligned}$$

The Cauchy-Schwarz inequality can be used to show that orthogonal projections are bounded operators, therefore

$$||\boldsymbol{SG}(\hat{\boldsymbol{y}}_{t+h|t} - \boldsymbol{y}_{t+h})||_2 \leq ||(\hat{\boldsymbol{y}}_{t+h|t} - \boldsymbol{y}_{t+h})||_2$$

Find reference or prove

□

The inequality is strict whenever $\hat{\boldsymbol{y}}_{t+h|t} \notin \mathfrak{s}$.

3.2 Probabilistic forecast reconciliation

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

Let $(\mathbb{R}^n, \mathscr{F}_{\mathbb{R}^n}, \hat{\nu})$ be an probability triple, that is incoherent and that characterises forecast uncertainty for all variables in the hierarchy at time $t+h$ conditional on all information up to time t. This may be obtained from the first stage of the forecasting process e.g. by modelling and forecasting each series individually. Let $g : \mathbb{R}^n \to \mathbb{R}^m$ be a linear mapping. Let $(\mathbb{R}^m, \mathscr{F}_{\mathbb{R}^m}, \nu)$ be a probability triple defined on $\mathbb{R}^m$.

Definition 3.2. We define the reconciled probability measure of $\hat{\nu}$ with respect to the mapping $g(.)$ as a probability measure $\tilde{\nu}$ on $\mathfrak{s}$ with σ-algebra $\mathscr{F}_{\mathfrak{s}}$ where the following holds

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

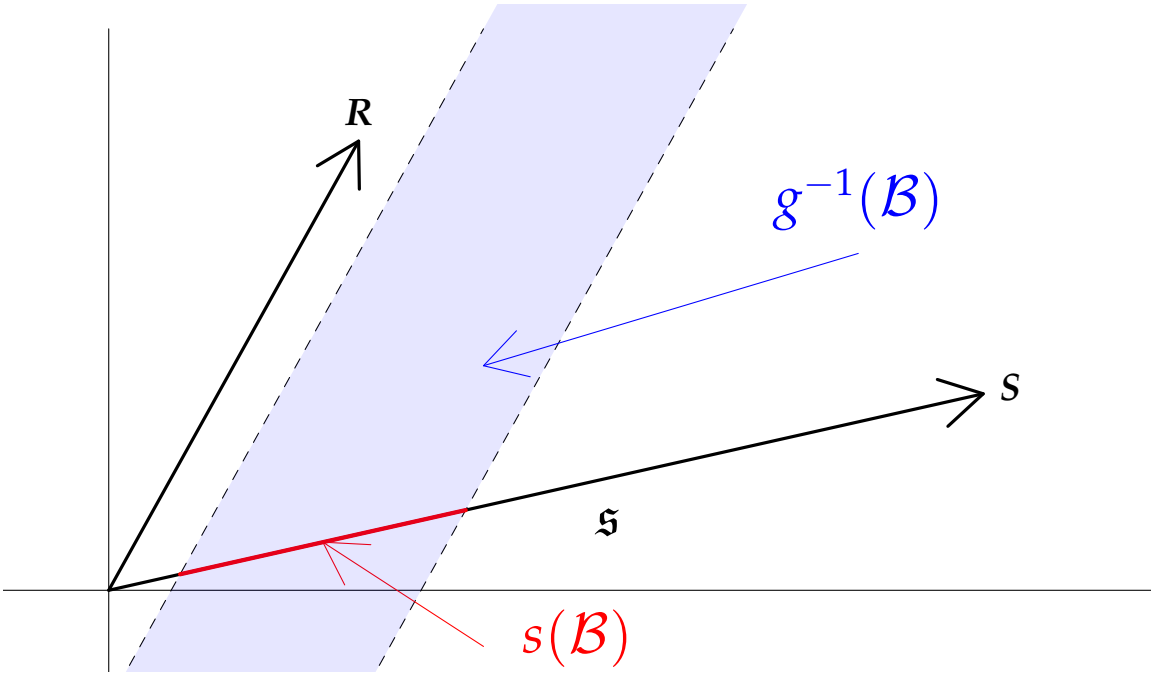

Figure 4: *Summary of probabilistic forecast reconciliation. The probability that $\boldsymbol{y}_{t+h|t}$ lies in the red line segment under the reconciled probabilistic forecast is defined to be equal to the probability that $\boldsymbol{y}_{t+h|t}$ lies in the shaded blue area under the unreconciled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

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

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

Recall that when $s \circ g$ is a projection, the case of point forecast reconciliation could be broken down into three steps. In the first, $\hat{\boldsymbol{y}}_{t+h|t}$ is transformed into coordinates $\tilde{\boldsymbol{b}}_{t+h|t}$ and $\tilde{\boldsymbol{a}}_{t+h|t}$ via

a change of basis. In the second, $\tilde{\boldsymbol{a}}_{t+h|t}$ is discarded and $\tilde{\boldsymbol{b}}_{t+h|t}$ are kept as the bottom level reconciled forecasts. In the third, reconciled forecasts for the entire hierarchy are recovered via $\tilde{\boldsymbol{y}}_{t+h|t} = \boldsymbol{S}\tilde{\boldsymbol{b}}_{t+h|t}$. We now outline the analogues to these three steps for probabilistic forecasts when predictive densities are available.

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

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

The expression $\hat{\nu}(g^{-1}(\mathcal{B}))$ in Definition 3.2 is equivalent to the probability statement $\Pr(\hat{\boldsymbol{y}}_{t+h|t} \in g^{-1}(\mathcal{B}))$. After the change of basis this is equivalent to $\Pr(\tilde{\boldsymbol{b}} \in \mathcal{B})$ which implies

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

After integrating out over $\tilde{\boldsymbol{a}}_{t+h|t}$, a step analogous to setting $\tilde{\boldsymbol{a}}_{t+h|t} = 0$ for point forecasting, we obtain an expression that gives the probability the reconciled bottom level series lies in the region $\mathcal{B}$. This corresponds to $\nu(\mathcal{B})$ in Definition 3.2. To make a valid probability statement about the entire hierarchy we simply use the bottom level probabilistic forecasts together with Definition 2.3.

Example: Gaussian Distributions

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

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

After a change in basis

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

where

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

The quadratic form q can be rearranged as

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

Recall that

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

Then q can be rearranged further as

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

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

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

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

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

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

3.3 Elliptical Distributions

We now show that the true predictive distribution can be recovered for elliptical distributions via linear reconciliation. Here, for any square matrix $\boldsymbol{C}$, $\boldsymbol{C}^{1/2}$ and $\boldsymbol{C}^{-1/2}$ are defined to satisfy $\boldsymbol{C}^{1/2}\left(\boldsymbol{C}^{1/2}\right)' = \boldsymbol{C}$ and $\boldsymbol{C}^{-1/2}\left(\boldsymbol{C}^{-1/2}\right)' = \boldsymbol{C}^{-1}$, for example $\boldsymbol{C}^{1/2}$ may be obtained via the Cholesky or eigenvalue decompositions.

Theorem 3.3 (Reconciliation for Elliptical Distributions)**.** *Let an unreconciled probabilistic forecast come from the elliptical class with location parameter $\hat{\boldsymbol{\mu}}$ and scale matrix $\hat{\boldsymbol{\Sigma}}$. Let the true predictive distribution of $\boldsymbol{y}_{t+h|t}$ also belong to the elliptical class with location parameter $\boldsymbol{\mu}$ and scale matrix $\boldsymbol{\Sigma}$. Then the affine reconciliation mapping $g(\breve{\boldsymbol{y}}) = \boldsymbol{G}_{opt}\breve{\boldsymbol{y}} + \boldsymbol{d}_{opt}$ with $\boldsymbol{G}_{opt} = \boldsymbol{A}\boldsymbol{\Sigma}^{-1/2}$ 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 the true variance covariance matrix of the predictive distribution for the bottom level.*

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

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

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

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

□

A number of insights can be drawn from this theorem. First, although a linear mapping $g(.)$ can be used to recover the true density 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 reconciliation are given below.

Theorem 3.4 (True predictive via projection)**.** *Assume that the true predictive distribution is elliptical with location* $\boldsymbol{m}u$ *and scale* $\boldsymbol{S}igma$*. 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* $\hat{\boldsymbol{\mu}}$ $\hat{\boldsymbol{\Sigma}}$ *and scale when the following conditions hold.*

$$\begin{aligned} sp(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) &\subset sp(\boldsymbol{R}) \\ sp(\boldsymbol{\Sigma}^{1/2} - \boldsymbol{\Sigma}^{1/2}) &\subset sp(\boldsymbol{R}) \end{aligned}$$

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 simplified to $\boldsymbol{\mu}$. If $\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}$ lies in the span of $\boldsymbol{R}$ then multiplication by $\boldsymbol{R}'_{\perp}$ reduced the second term to $\mathbf{0}$. 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. As such an orthogonal projection after transformation by the inverse of the one step ahead in-sample errors may be more intuitively appealing. This is exactly what the MinT projection estimates, and as simulations will show in Section 5, this projection leads to the best empirical results.

4 Evaluation of hierarchical probabilistic forecasts

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

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

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

Let 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

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

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

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

Some References

4.1 Univariate Scoring rules

One way to evaluate probabilistic forecasts is via the application of univariate scoring rules to each variable in a hierarchy. A summary can be taken of the expected scores across each margin for example a mean or median. In the simulations of section 5 we consider two scoring rules. The log score is given by the log of the marginal density of each variable. The cumulative rank probability score generalises mean square error and is given by

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

where $\breve{Y}_i$ and $\breve{Y}^{*}_{T+h,i}$ are independent copies of a random variable with distribution $\breve{F}_i$ and the latter is the predictive distribution for the i^{th} margin of a forecast for time $t+h$ made at time t. The expectations in the second line can be approximated by Monte Carlo when a sample from the predictive distribution is available.

An advantage to this approach is that it allows the forecaster to evaluate the levels and individual series of the hierarchy where the gains to reconciliation are greatest. For this reason this approach

has been used in the limited literature on probabilistic forecasting for hierarchies Ben Taieb et al., 2017 and Jeon et al to date. A major shortcomings to this approach however is that, evaluating univariate scores on the margins do not account for the dependence in the hierarchy.

Jeon Panagiotelis and Petropoulos reference

4.2 Multivariate Scoring rules

While the a number of alternative proper scoring rules are available for univariate forecasts, the multivariate case is somewhat more limited. Here we focus on three scoring rules: the log score, the energy score and the variogram score. These are summarized in table 2.

Table 2: *Scoring rules to evaluate multivariate forecast densities. Here, $\breve{\boldsymbol{y}}_{T+h}$ and $\breve{\boldsymbol{y}}^*_{T+h}$ are two independent random vectors from the coherent forecast distribution $\breve{\boldsymbol{F}}$ with density function $\breve{\boldsymbol{f}}(\cdot)$ at time $T+h$, and $\boldsymbol{y}_{T+h}$ is the vector of realizations. Further, $\breve{Y}_{T+h,i}$ and $\breve{Y}_{T+h,j}$ are the ith and jth components of the vector $\breve{\boldsymbol{Y}}_{T+h}$. The variogram score is given for order p, where w_{ij} denote non-negative weights.*

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

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

The energy score on the other hand can be defined in terms of the characteristic function of the probabilistic forecast, but the representation in Table 2 in terms of expectations leads itself to easy computation when samples from the probabilistic forecast are available. An interesting case is where $\alpha = 2$, where it can be easily shown that

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

where $\breve{\boldsymbol{\mu}}_{T+h} = \mathrm{E}_{\boldsymbol{F}}(\breve{\boldsymbol{Y}}_{T+h})$. In this limiting case, the energy score only measures the accuracy of the forecast mean, and not the entire distribution and the energy score is proper, but not strictly proper. Pinson and Tastu (2013) also argues that the energy score has very low discriminative ability for incorrectly specified covariances, even though it discriminates the misspecified means well.

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

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

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

4.2.1 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 2. As such it may seem sensible to compare unreconciled densities to reconciled densities on the basis of log score. The following theorem shows that using the log score may fail in the case of multivariate distributions with a degeneracy.

Theorem 4.1 (Impropriety of log score)**.** *When the true data generating process is a coherent measure, 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 $n \times n$ $\boldsymbol{U}$ could be the matrix of left singular vectors of $\boldsymbol{S}$. For a non-degenerate probability measure on $\mathbb{R}^n$, the density is the Radon-Nikodym derivative with respect to the usual Lesbegue measure on $\mathbb{R}^n$.

Consider the case where the true density is $f_1(\boldsymbol{z}_t^{(1)})\mathbb{1}\left(\boldsymbol{z}_t^{(2)}\right)$, and is compared to an incoherent density is given by $f_1(\boldsymbol{z}_t^{(1)})f_2(\boldsymbol{z}_t^{(2)})$, where f_2 is highly concentrated around 0 but still a proper density. 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 DGP is

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

while that of the unreconciled density is

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

After taking expectations $ES(f, f) > ES(\hat{f}, f)$, violating the condition (1) for a proper scoring rule. □

A similar issue also arises when discrete random variables are modelled as if they were continuous, an issue discussed in Section 4.1, page 366 of Gneiting and Raftery, 2007. This implies that the log score should not be used to evaluate multivariate densities with degeneracies and should be avoided when comparing reconciled and unreconciled probabilistic forecasts.

4.2.2 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 2.3. 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 is dependent 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 when assessed using the bottom level series, 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, 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 and Rizzo, 2013; Gneiting and Raftery, 2007). In the context of two coherent forecasts the same

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

Table 3: *Summary of properties of scoring rules in the context of reconciled probabilistic forecasts.*

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. As such 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 energy score using the full hierarchy. Although the discussion here is related to energy score, the same logic holds for other multivariate scores that are not invariant to orthogonal rotations, for example the variogram score.

The properties of multivariate scoring rules in the context of evaluating reconciled probabilistic forecasts in Table 3.

5 Simulation Study

We now turn our attention to comparing different reconciliation methods in a simulation study where the data is conditionally Gaussian. We choose the Gaussian case due to its analytical tractability which allows for evaluation on the basis of all scoring rules (including the log score). The non-Gaussian case lies beyond the scope of this simulation study, but can be handled by a bootstrapping approach proposed in separate work.

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

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 errors driving these ARIMA processes denoted ε are jointly normal $\{\varepsilon_{AA,t}, \varepsilon_{AB,t}, \varepsilon_{BA,t}, \varepsilon_{BB,t}\} \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})\ \forall t$. The parameters for the AR and MA components are randomly and uniformly generated from

$[0.3, 0.5]$ and $[0.3, 0.7]$ respectively. Then the bottom-level series $\{y_{AA,t}, y_{AB,t}, y_{BA,t}, y_{BB,t}\}$ are given by:

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

where $u_t \sim \mathcal{N}(0, \sigma_u^2)$ and $v_t \sim \mathcal{N}(0, \sigma_v^2)$. The aggregate series at in the middle level level, are given by:

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

and the total series is given by

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

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

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

and similar inequalities hold when $\varepsilon_{AA,t}$ is replaced by $\varepsilon_{AB,t}$, $\varepsilon_{BA,t}$ and $\varepsilon_{BB,t}$ in the second and third terms. The values of $\mathbf{\Sigma}$, σ_u^2 and σ_v^2, that we use and which satisfy these constraints are:

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

We generate data a sample size of $T = 501$. Univariate ARIMA models are selected for each series using the *auto.arima* function in the *forecast* package (Hyndman, 2017) in R (R Core Team, 2018). The same package was used to fit each series independently using the first 500 observations, and evaluate 1-step ahead base (incoherent) probabilstic forecasts. These were then reconciled using different projections summarzied in Table 1. This process was replicated using 1000 different data sets from the same data generating processes.

Are the DGPs the same in all cases?

Table 4: *Comparison of coherent forecasts. "Energy score" and "Variogram score" columns give scores based on the joint forecast distribution of whole hierarchy. "Log score" column gives the log scores of the joint forecast distribution of bottom level. "Skill score" columns give the percentage skill score with reference to the bottom-up method. Entries in these columns show the percentage increase of score for different reconciliation methods relative to the bottom-up method.*

Forecasting	Energy score		Variogram score		Log score	
method	Mean score	Skill score %	Mean score	Skill score %	Mean score	Skill score %
MinT(Shrink)	10.03	18.79	8.44	8.46	11.30	6.22
MinT(Sample)	10.01	18.95	8.41	8.79	11.29	6.31
MinT(WLS)	10.53	14.74	9.02	2.17	12.61	−4.65
OLS	10.53	14.74	8.86	3.09	11.54	4.23
Bottom-up	12.35		9.22		12.05	
Incoherent	11.12		9.53			

To assess the predictive performance of different forecasting methods, we use scoring rules as discussed in Section 4. To facilitate comparisons, we report skill scores (Gneiting and Raftery, 2007). For a given forecasting method, evaluated by a particular scoring rule $S(\cdot)$, the skill score gives the percentage improvement of the preferred forecasting method relative to a reference method. A negative valued skill score indicates that a method is worse than the reference method, whereas any positive value indicates that method is superior to the reference method.

Table 4 summarizes the forecasting performance of unreconciled, bottom-up, OLS, WLS and two MinT reconciliation methods using log score, energy score and variogram score. In all cases skill scores are calculated with the bottom-up method as reference. 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. The cell for log score of unreconciled forecasts is left blank since the log score is not proper in this context. 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 unreconciled forecasts. Bottom up forecasts perform even worse than unreconciled forecasts.

Tables 5 and 6 break down the forecasting performance of different reconciliation methods by considering univariate scores on each individual margin. The log score and CRPS are considered while skill scores are computed with the unreconciled forecast as a reference. When broken down in this fashion, the methods based on MinT perform best for all series and always outperform bottom up and unreconciled forecasts. The same cannot be said for OLS which performs worse than bottom up and incoherent forecasts on every individual series.

Puzzling it does so bad, but outperforms unrec-

Table 5: *Comparison of incoherent vs coherent forecasts based on the univariate forecast distribution of aggregate series. The "Incoherent" row shows the average scores for incoherent forecasts. Each entry above this row represents the percentage skill score with reference to the incoherent forecasts. These entries show the percentage increase in score for different forecasting methods relative to the incoherent forecasts.*

Forecasting	Total		Series - A		Series - B	
method	CRPS	LogS	CRPS	LogS	CRPS	LogS
MinT(Shrink)	0.74	0.00	10.49	3.24	9.16	2.73
MinT(Sample)	0.74	0.00	10.49	3.24	9.16	2.73
MinT(WLS)	−2.96	−2.36	6.10	−4.12	5.66	−3.03
OLS	−9.26	−3.36	7.07	2.06	7.01	1.82
Bottom-up	−91.48	−22.22	−8.05	−2.06	−6.20	−1.82
Incoherent	*2.70*	*2.97*	*4.10*	*3.40*	*3.71*	*3.30*

Table 6: *Comparison of incoherent vs coherent forecasts based univariate forecast distribution of bottom-level series. The "Incoherent" row shows the average scores for incoherent forecasts.*

Forecasting	Series - AA		Series - AB		Series - BA		Series - BB	
method	CRPS	LogS	CRPS	LogS	CRPS	LogS	CRPS	LogS
MinT(Shrink)	7.61	2.43	10.82	3.02	5.93	1.86	7.76	2.47
MinT(Sample)	7.88	2.43	11.08	3.02	6.20	1.86	8.05	2.47
MinT(WLS)	3.53	0.00	6.33	0.60	2.43	−0.62	4.89	0.62
OLS	2.99	0.91	5.28	1.51	2.90	0.62	4.31	1.23
Incoherent	*3.68*	*3.29*	*3.79*	*3.31*	*3.45*	*3.22*	*3.48*	*3.24*

6 Conclusions

By redefining coherent forecasts and forecast reconciliation in geometric terms we have established two new theoretical results that support the use of projections for point forecast reconciliation, and crucially have extended these concepts to probabilistic forecasting. We show that for elliptical distributions, the true predictive density can be recovered by linear reconciliation and establish conditions for which this linear mapping is a projection. Although this optimal projection cannot feasibly be obtained in practice, a projection similar to the MinT procedure provides a good approximation in practice. This is supported by the results of a simulation study. Finally, we also discuss strategies for evaluating probabilistic forecasts for hierarchical time series establishing a key results regarding the impropriety of the log score with respect to incoherent forecasts.

In many ways this paper sets up a substantial future research agenda. For example, having defined what amounts to an entire class of reconciliation methods for probabilistic forecasts it will be worthwhile investigating which specific projections are optimal. This is likely to

depend on the specific scoring rule employed as well as the properties of base forecasts. Another avenue worth investigation is to consider whether it is possible to recover the true predictive distribution for non-elliptical distributions possible via a non-linear mapping $g(.)$.

References

Athanasopoulos, G, Hyndman, RJ, Kourentzes, N, and Petropoulos, F (2017). Forecasting with temporal hierarchies. *European Journal of Operational Research* **262**(1), 60–74.

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

Dunn, DM, Williams, WH, and Dechaine, TL (1976). Aggregate Versus Subaggregate Models in Local Area Forecasting. *Journal of American Statistical Association* **71**(353), 68–71.

Erven, T van and Cugliari, J (2014). *Game-Theoretically Optimal reconciliation of contemporaneous hierarchical time series forecasts*. Ed. by A Antoniadis, X Brossat, and J Poggi, pp. 297–317.

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

Gneiting, T and Raftery, AE (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association* **102**(477), 359–378.

Gneiting, T, Stanberry, LI, Grimit, EP, Held, L, and Johnson, NA (2008). "Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds".

Gross, CW and Sohl, JE (1990). Disaggregation methods to expedite product line forecasting. *Journal of Forecasting* **9**(3), 233–254.

Hyndman, R (2017). forecast: Forecasting Functions for Time Series and Linear Models, R package version 8.0. *URL: http://github.com/robjhyndman/forecast*.

Hyndman, RJ, Ahmed, RA, Athanasopoulos, G, and Shang, HL (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics and Data Analysis* **55**(9), 2579–2589.

Jordan, A, Krüger, F, and Lerch, S (2017). Evaluating probabilistic forecasts with the R package scoringRules. arXiv: 1709.04743.

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

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

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

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

Székely, GJ and Rizzo, ML (2013). Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference* **143**(8), 1249–1272.

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