# Probabilistic Reconciliation of Hierarchical Forecast via Bayes' Rule

Giorgio Corani, Dario Azzimonti, João P. S. C. Augusto, and Marco Zaffalon

Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA) Manno, Switzerland
giorgio{dario.azzimonti,zaffalon}@idsia.ch

**Abstract.** We present a novel approach for reconciling hierarchical forecasts, based on Bayes' rule. We define a prior distribution for the bottom time series of the hierarchy, based on the bottom base forecasts. Then we update their distribution via Bayes rule, based on the base forecasts for the upper time series. Under the Gaussian assumption, we derive the updating in closed-form. We derive two algorithms, which differ as for the assumed independencies. We discuss their relation with the MinT reconciliation algorithm and with the Kalman filter, and we compare them experimentally.

## 1 Introduction

Often time series are organized into a hierarchy. For example, the total visitors of a country can be divided into regions and the visitors of each region can be further divided into sub-regions. The most disaggregated time series of the hierarchy are referred to as *bottom time series*, while the remaining time series are referred to as *upper time series*.

Forecasts of hierarchical time series should be *coherent*; for instance, the sum of the forecasts of the different regions should equal the forecast for the total. The forecasts are *incoherent* if they do not satisfy such constraints. A simple way for generating coherent forecasts is *bottom-up*: one takes the forecasts for the bottom time series and sums them up according to the summing constraints in order to produce forecasts for the entire hierarchy. Yet this approach does not consider the forecasts produced for the upper time series, which contain useful information. For instance, upper time series are smoother and allow to better estimate of the seasonal patterns and the effect of the covariates.

Thus, modern reconciliation methods [9, 18] proceed in two steps. First, *base forecasts* are computed by fitting an independent model to each time series. Then, the base forecasts are adjusted to become coherent; this step is called *reconciliation*. The forecasts for the entire hierarchy are then obtained by summing up the reconciled bottom time series. Reconciled forecasts are generally more accurate than the base forecasts, as they benefit from information coming from multiple time series. The state-of-the art reconciliation algorithm is MinT [18], which minimizes the mean squared error of the reconciled forecasts by solving a generalized least squares problem; its point forecast, besides being coherent, are generally more accurate than the base forecast.

Hierarchical probabilistic forecasting is however still an open area of research. The algorithm by [17] constructs a coherent forecast in a bottom-up fashion, modelling via copulas the joint distribution of the bottom time series, while [13] proposes a top-down approach, where the top time series is forecasted and then disaggregated. Both algorithms are based on numerical procedures which have no closed-form solution; hence they are not easily interpretable. In [6], a geometric interpretation of the reconciliation process is provided. It is moreover shown that the log score is not proper with respect to incoherent probabilistic forecasts. As an alternative, the energy score can be used for comparing reconciled to unreconciled probabilistic hierarchical forecasts. In [1] multivariate Gaussian predictive densities and bootstrap densities are experimentally compared for hierarchical probabilistic forecasting.

We address probabilistic reconciliation using Bayes' rule. We define the prior beliefs about the bottom time series, based on the base forecasts for the bottom time series. We then update them incorporating the information contained in the forecasts for the upper time series. Under the Gaussian assumption, we compute the update in closed form, obtaining the posterior distribution about the bottom time series and then about the entire hierarchy. Our reconciled forecasts minimize the mean squared error; indeed, we prove that they match the point predictions of MinT, whose optimality has been proven in a frequentist way. Our algorithm provides the joint predictive distribution for the hierarchy; thus we call it pMinT, which stands for probabilistic MinT. We also provide a variant of pMinT, obtained by making an additional independence assumption; we call it LG, as it is related to the linear-Gaussian model [3, Chap. 8.1.4].

We show a link between the reconciliation problem and the Kalman filter, opening the possibility of borrowing from the literature of the Kalman filter for future research. We then compare the algorithms on synthetic and real data sets, eventually concluding that pMinT yields more accurate probabilistic forecasts than both bottom-up and LG.

The paper is organized as follows: we introduce the reconciliation problem in Section 2, we discuss the algorithms in Sec. 3, we discuss the reconciliation of a simple hierarchy in Section 4 and we present the experiments in Section 5.

## 2 Time series reconciliation

Fig. 1 shows a hierarchy. We could interpret it as the visitors of a country, which are disaggregated first by region ($R_1$, $R_2$) and then by sub-regions ($R_{11}$, $R_{12}$, $R_{21}$, $R_{22}$). The most disaggregated time series (*bottom time series*) are shaded. The hierarchy contains $m$ time series, of which $n$ are bottom time series. We denote by uppercase letters the random variables and by lowercase letters their observations. The vector of observations available at time $t$ for the entire hierarchy is $\mathbf{y}_t \in \mathbb{R}^m$; they are observations from the set of random variables $\mathbf{Y}_t$. Vector $\mathbf{y}_t$ can be broken down in two parts, namely $\mathbf{y}_t = [\mathbf{u}_t^T, \mathbf{b}_t^T]^T$; $\mathbf{b}_t \in \mathbb{R}^n$ contains the observations of the bottom time series while $\mathbf{u}_t \in \mathbb{R}^{m-n}$ contains the observations of the upper time series. At time $t$, the observations available for the hierarchy

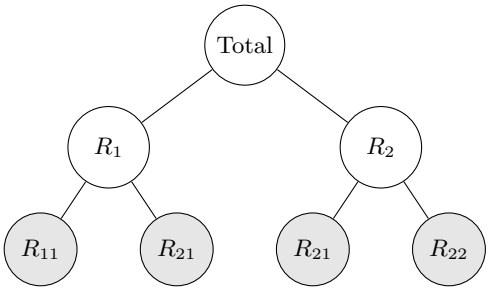



Fig. 1: A hierarchical time series which disaggregates the visitors into regions and sub-regions.

of Fig. 1 are thus:

$$\mathbf{y}_t = [y_{\text{Total}}, y_{R_1}, y_{R_2}, y_{R_{11}}, y_{R_{12}}, y_{R_{21}}, y_{R_{22}}]^T = [\mathbf{u}_t^T, \mathbf{b}_t^T]^T,$$

where:

$$\mathbf{u}_t = [y_{\text{Total}}, y_{R_1}, y_{R_2}]^T$$

$$\mathbf{b}_t = [y_{R_{11}}, y_{R_{12}}, y_{R_{21}}, y_{R_{22}}]^T.$$

The structure of the hierarchy is represented by the summing matrix $\mathbf{S} \in \mathbb{R}^{m \times n}$ such that:

$$\mathbf{y}_t = \mathbf{S}\mathbf{b}_t. \tag{1}$$

The $\mathbf{S}$ matrix of hierarchy in Fig.1 is:

$$\mathbf{S} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \mathrm{A} \\ \hline \mathrm{I} \end{bmatrix}, \tag{2}$$

where the sub-matrix $\mathbf{A} \in \mathbb{R}^{(m-n) \times n}$ encodes which bottom time series should be summed up in order to obtain each upper time series.

We denote by $\widehat{\mathbf{y}}_{t+h} \in \mathbb{R}^m$ the base forecasts issued at time $t$ about of $y$ and referring to $h$ steps ahead. We separate base forecasts for bottom time series ($\widehat{\mathbf{b}}_{t+h} \in \mathbb{R}^n$) and upper time series ($\widehat{\mathbf{u}}_{t+h} \in \mathbb{R}^{m-n}$), namely $\widehat{\mathbf{y}}_{t+h} = [\widehat{\mathbf{u}}_{t+h}^T, \widehat{\mathbf{b}}_{t+h}^T]^T$. The variances of the error of the base forecasts will be used later. If forecasts for different time horizons are needed (e.g., $h$=1,2,3,..), the reconciliation is performed independently for each $h$. In the following we generically assume to reconcile the forecasts for $h$ steps ahead.

*The MinT reconciliation* Most reconciliation algorithms [9], including MinT [18], assume the reconciled bottom forecasts ($\widetilde{\mathbf{b}}_{t+h}$) to be a linear combination of the base forecasts ($\widehat{\mathbf{y}}_{t+h}$) available for the whole hierarchy, i.e. their objective is to find a matrix $\mathbf{P}_h \in \mathbb{R}^{n \times m}$ such that:

$$\widetilde{\mathbf{b}}_{t+h} = \mathbf{P}_h \widehat{\mathbf{y}}_{t+h}. \tag{3}$$

Let us denote by $\widehat{\mathbf{E}}_{t+h} = \mathbf{Y}_{t+h} - \widehat{\mathbf{y}}_{t+h} \in \mathbb{R}^m$ the vector of the errors of the base forecast $h$-steps ahead and by $\mathbf{W}_h = \mathbb{E}[\widehat{\mathbf{E}}_{t+h}\ \widehat{\mathbf{E}}_{t+h}^T \mid \mathcal{I}_t]$ their covariance matrix, where $\mathcal{I}_t$ denotes all the information available up to time $t$. In [18] it is proven that the reconciliation matrix given by:

$$\mathbf{P}_h = (\mathbf{S}^T \mathbf{W}_h^{-1} \mathbf{S})^{-1} \mathbf{S}^T \mathbf{W}_h^{-1} \tag{4}$$

is optimal, in the sense that it minimizes the trace of the reconciliation errors' covariance matrix. The reconciled forecasts for the whole hierarchy are obtained by summing the reconciled bottom forecasts, and they are proven to minimize the mean squared error over the entire hierarchy.

**Estimation of $\mathbf{W}_h$** Estimating $\mathbf{W}_h$ differently for each $h$ is an open problem. For the case $h$=1 the estimation is simpler. The variance of the forecasts equals the variance of the residuals (i.e., the errors on 1-step predictions made on the training data) and cross-covariances are estimated as the covariance of the residuals. The best estimates are obtained [18] by shrinking the full covariance matrix towards a diagonal matrix, using the method of [15].

The case $h$ >1 is instead problematic. The variance of the forecasts are obtained by increasing the 1-step variance through analytical formulas, which differ from the variance of the $h$-steps ahead residuals. Morevoer, the covariances in $\mathbf{W}_h$ have to be numerically estimated by looking at the $h$-steps residuals. However, the number of $h$-steps residuals decreases with $h$, making the estimate more noisy.

As a workaround, [18] assumes $\mathbf{W}_h = k_h \mathbf{W}_1$, where $k_h > 0$ is an unknown constant which depends on $h$ while $\mathbf{W}_1$ is the covariance matrix of the one-step ahead errors. The underlying assumption is thus that all terms within the variance/covariance matrix of the errors grow in the same way with $h$. The advantage of this approach is that $k_h$ cancels out when computing the reconciled forecasts, as it can be seen by setting $\mathbf{W}_h = k_h \mathbf{W}_1$ in Eq.(4). Yet, $k_h$ appears in the expression of the variance of the reconciled forecasts. In the following we refer to the assumption $\mathbf{W}_h = k_h \mathbf{W}_1$ as "the $k_h$ assumption".

## 3 Probabilistic Reconciliation

We address the reconciliation problem by merging the probabilistic information contained in the base forecasts for the bottom and the upper time series. We perform the fusion using Bayes' rule.

We first define the prior about the *bottom* time series. We have observed the time series up to time $t$ and we are interested in the reconciled forecasts for time $t+h$. We denote by $\mathbf{B}_{t+h}$ the vector of the bottom time series at time $t+h$; this is thus a vector of random variables and $B^i_{t+h}$ represents its $i$th element. We moreover denote by $\widehat{\mathbf{b}}_{t+h}$ the vector of base forecasts for the bottom time series for time $t+h$, and by $\mathbf{b}_{t+h}$ the actual observation of the bottom random variables at time $t+h$. Finally, $\mathcal{I}_{t,b}$ is the information available up to time $t$ regarding the bottom time series, i.e. the past values of the bottom time series: $\mathcal{I}_{t,b} = \{\mathbf{b}_1, \dots, \mathbf{b}_t\}$.

In the following we adopt the $k_h$ assumption for all the covariance matrices assuming moreover that, for a given $h$, the value of $k_h$ is shared among all the involved covariance matrices. As we will show later, this is equivalent to the $k_h$ assumption made by MinT. Let us hence denote the covariance matrix of the forecast $h$-steps ahead by $\widehat{\boldsymbol{\Sigma}}_{B,h} = k_h \widehat{\boldsymbol{\Sigma}}_{B,1}$. Assuming the bottom time series to be jointly Gaussian we have:

$$p(\mathbf{B}_{t+h} \mid \mathcal{I}_{t,b}) = N\left(\widehat{\mathbf{b}}_{t+h}, k_h \widehat{\boldsymbol{\Sigma}}_{B,1}\right). \tag{5}$$

*Probabilistic bottom-up* If we have no information about the upper time series, we can build a joint predictive distribution for the entire hierarchy by summing the bottom forecast via matrix $\mathbf{S}$:

$$p(\mathbf{Y}_{t+h} \mid \mathcal{I}_{t,b}) = N\left(\mathbf{S}\widehat{\mathbf{b}}_{t+h}, \mathbf{S} k_h \widehat{\boldsymbol{\Sigma}}_{B,1} \mathbf{S}^T\right), \tag{6}$$

which is a *probabilistic bottom-up* reconciliation. Note that, in this case, $k_h$ appears only in the expression of the variance.

*Updating* If the forecasts $\widehat{\mathbf{U}}_{t+h}$ about the upper time series are available, then we can use them in order to update our prior. We assume:

$$\begin{aligned} &\widehat{\mathbf{U}}_{t+h} = \mathbf{A}\mathbf{B}_{t+h} + \boldsymbol{\varepsilon}^u_{t+h}, \\ &\boldsymbol{\varepsilon}^u_{t+h} \sim N\left(\mathbf{0}, \widehat{\boldsymbol{\Sigma}}_{U,h}\right), \end{aligned} \tag{7}$$

where $\widehat{\boldsymbol{\Sigma}}_{U,h} = k_h \widehat{\boldsymbol{\Sigma}}_{U,1}$ is the covariance of the noise. We thus treat $\widehat{\mathbf{U}}_{t+h}$ as a set of different sums of the future values of the bottom time series, corrupted by noise. Hence:

$$p(\widehat{\mathbf{U}}_{t+h} \mid \mathbf{B}_{t+h}) = N\left(\mathbf{A}\mathbf{B}_{t+h}, k_h \widehat{\boldsymbol{\Sigma}}_{U,1}\right). \tag{8}$$

The posterior distribution of the bottom time series is given by Bayes' rule:

$$\begin{aligned} p(\mathbf{B}_{t+h} \mid \mathcal{I}_{t,b}, \widehat{\mathbf{U}}_{t+h}) &= \frac{p(\mathbf{B}_{t+h} \mid \mathcal{I}_{t,b}) p(\widehat{\mathbf{U}}_{t+h} \mid \mathcal{I}_{t,b}, \mathbf{B}_{t+h})}{p(\widehat{\mathbf{U}}_{t+h} \mid \mathcal{I}_{t,b})} = \\ &= \frac{p(\mathbf{B}_{t+h} \mid \mathcal{I}_{t,b}) p(\widehat{\mathbf{U}}_{t+h} | \mathbf{B}_{t+h})}{p(\widehat{\mathbf{U}}_{t+h} \mid \mathcal{I}_{t,b})} \propto \\ &\propto p(\mathbf{B}_{t+h} \mid \mathcal{I}_{t,b}) p(\widehat{\mathbf{U}}_{t+h} \mid \mathbf{B}_{t+h}) = \\ &= p(\mathbf{B}_{t+h} \mid \mathcal{I}_{t,b}) p(\mathbf{A}\mathbf{B}_{t+h} + \boldsymbol{\varepsilon}^u_{t+h} \mid \mathbf{B}_{t+h}) \end{aligned} \tag{9}$$

### 3.1 Computing Bayes' rule

The posterior of Eq. (9) can be computed in closed form by assuming the vector $(\mathbf{B}_{t+h}, \widehat{\mathbf{U}}_{t+h})$ to be jointly Gaussian distributed. The linear-Gaussian (LG) model [3, Chap. 8.1.4]. computes analytically the updating by further assuming $\boldsymbol{\varepsilon}^u_{t+h}$ to be independent from $\mathbf{B}_{t+h}$. Yet this independence might not always hold in our case. Consider for instance a special event driving upwards most time series. As a result we would observe both high values of $\mathbf{B}_{t+h}$ and negative values of $\boldsymbol{\varepsilon}^u_{t+h}$, due to the underestimation of the upper time series. This would result in a correlation between $\mathbf{B}_{t+h}$ and $\boldsymbol{\varepsilon}^u_{t+h}$.

We thus generalize the LG model by accounting for such correlation. We will later compare experimentally the results obtained adopting the LG model and its generalized version. We denote $\mathrm{Cov}(\mathbf{B}_{\mathrm{t+1}}, \boldsymbol{\varepsilon}^{\mathrm{u}}_{\mathrm{t+1}} \mid \mathcal{I}_{\mathrm{t,b}}) = \mathbf{M}_1 \in \mathbb{R}^{n\times(m-n)}$ and we assume $\mathrm{Cov}(\mathbf{B}_{\mathrm{t+h}}, \boldsymbol{\varepsilon}^{\mathrm{u}}_{\mathrm{t+h}} \mid \mathcal{I}_{\mathrm{t,b}}) = k_h\mathbf{M}_1$.

Our first step for computing Bayes' rule is to express the joint distribution $p(\mathbf{B}_{t+h}, \widehat{\mathbf{U}}_{t+h} \mid \mathcal{I}_{t,b})$. Since $\widehat{\mathbf{U}}_{t+h} = \mathbf{A}\mathbf{B}_{t+h} + \boldsymbol{\varepsilon}^u_{t+h}$, the expected values are:

$$\begin{aligned}\mathbb{E}[\mathbf{B}_{t+h} \mid \mathcal{I}_{t,b}] &= \widehat{\mathbf{b}}_{t+h},\\ \mathbb{E}[\widehat{\mathbf{U}}_{t+h} \mid \mathcal{I}_{t,b}] &= \mathbf{A}\widehat{\mathbf{b}}_{t+h}.\end{aligned}$$

We now derive the different blocks of the covariance matrix. The cross-covariance between $\mathbf{B}_{t+h}$ and $\widehat{\mathbf{U}}_{t+h}$ is:

$$\begin{aligned}\mathrm{Cov}(\mathbf{B}_{\mathrm{t+h}}, \widehat{\mathbf{U}}_{\mathrm{t+h}} \mid \mathcal{I}_{\mathrm{t,b}}) &= \mathrm{Cov}(\mathbf{B}_{\mathrm{t+h}}, \mathbf{A}\mathbf{B}_{\mathrm{t+h}} + \boldsymbol{\varepsilon}^{\mathrm{u}}_{\mathrm{t+h}} \mid \mathcal{I}_{\mathrm{t,b}})\\ &= \mathrm{Cov}(\mathbf{B}_{\mathrm{t+h}}, \mathbf{B}_{\mathrm{t+h}} \mid \mathcal{I}_{\mathrm{t,b}})\,\mathbf{A}^T + \mathrm{Cov}(\mathbf{B}_{\mathrm{t+h}}, \boldsymbol{\varepsilon}^{\mathrm{u}}_{\mathrm{t+h}} \mid \mathcal{I}_{\mathrm{t,b}})\\ &= k_h(\widehat{\boldsymbol{\Sigma}}_{B,1}\mathbf{A}^T + \boldsymbol{M}_1) \in \mathbb{R}^{n\times(m-n)}\end{aligned}$$

where $k_h > 0$ is the multiplicative constant (Sec.2) that yields the variance of the forecasts $h$-steps ahead, given the covariances of the forecasts 1-step ahead.

The covariance of the upper forecasts is:

$$\begin{aligned}\mathrm{Cov}(\widehat{\mathbf{U}}_{t+h} \mid \mathcal{I}_{t,b}) &= \mathrm{Cov}(\mathbf{A}\mathbf{B}_{t+h} + \boldsymbol{\varepsilon}^u_{t+h}, \mathbf{A}\mathbf{B}_{t+h} + \boldsymbol{\varepsilon}^u_{t+h} \mid \mathcal{I}_{t,b})\\ &= k_h\mathbf{A}\widehat{\boldsymbol{\Sigma}}_{B,1}\mathbf{A}^T + k_h\widehat{\boldsymbol{\Sigma}}_{U,1} + \mathrm{Cov}(\mathbf{A}\mathbf{B}_{\mathrm{t+h}}, \boldsymbol{\varepsilon}^{\mathrm{u}}_{\mathrm{t+h}}) + \mathrm{Cov}(\boldsymbol{\varepsilon}^{\mathrm{u}}_{\mathrm{t+h}}, \mathbf{A}\mathbf{B}_{\mathrm{t+h}})\\ &= k_h(\mathbf{A}\widehat{\boldsymbol{\Sigma}}_{B,1}\mathbf{A}^T + \widehat{\boldsymbol{\Sigma}}_{U,1} + \mathbf{A}\mathbf{M}_1 + \mathbf{M}_1^T\mathbf{A}^T).\end{aligned}$$

Hence the joint prior (i.e., before observing $\widehat{\mathbf{U}}_{t+h}$) is:

$$\begin{pmatrix}\mathbf{B}_{t+h}\\ \widehat{\mathbf{U}}_{t+h}\end{pmatrix} \mid \mathcal{I}_{t,b} \sim N\left[\begin{pmatrix}\widehat{\mathbf{b}}_{t+h}\\ \mathbf{A}\widehat{\mathbf{b}}_{t+h}\end{pmatrix}, \begin{pmatrix}k_h\widehat{\boldsymbol{\Sigma}}_{B,1} & k_h(\widehat{\boldsymbol{\Sigma}}_{B,1}\mathbf{A}^T + \mathbf{M}_1)\\ k_h(\mathbf{A}\widehat{\boldsymbol{\Sigma}}_{B,1} + \mathbf{M}_1^T) & k_h(\mathbf{A}\widehat{\boldsymbol{\Sigma}}_{B,1}\mathbf{A}^T + \widehat{\boldsymbol{\Sigma}}_{U,1} + \mathbf{A}\mathbf{M}_1 + \mathbf{M}_1^T\mathbf{A}^T)\end{pmatrix}\right].$$

Now we receive the forecast $\widehat{\mathbf{u}}_{t+h}$ for the upper time series (recall that $\widehat{\mathbf{U}}_{t+h}$ denotes the random variables while $\widehat{\mathbf{u}}_{t+h}$ denotes observations). We obtain the posterior distribution for the bottom time series $P(\mathbf{B}_{t+h}|\widehat{\mathbf{u}}_{t+h})$ by applying the

standard formulas for the conditional distribution of a MVN distribution [12, Sec.4.3.1]. To have a shorter notation, let us define:

$$\begin{aligned}\mathbf{G} &= \left(k_h(\widehat{\boldsymbol{\Sigma}}_{B,1}\mathbf{A}^T + \mathbf{M}_1)\right)\left(k_h(\mathbf{A}\widehat{\boldsymbol{\Sigma}}_{B,1}\mathbf{A}^T + \widehat{\boldsymbol{\Sigma}}_{U,1} + \mathbf{A}\mathbf{M}_1 + \mathbf{M}_1^T\mathbf{A}^T)\right)^{-1} \\ &= \left(\widehat{\boldsymbol{\Sigma}}_{B,1}\mathbf{A}^T + \mathbf{M}_1\right)\left(\mathbf{A}\widehat{\boldsymbol{\Sigma}}_{B,1}\mathbf{A}^T + \widehat{\boldsymbol{\Sigma}}_{U,1} + \mathbf{A}\mathbf{M}_1 + \mathbf{M}_1^T\mathbf{A}^T\right)^{-1}, \end{aligned} \tag{10}$$

where $k_h$ disappears from the expression of $\mathbf{G}$.

The reconciled bottom time series have then the following mean and variance:

$$\widetilde{\mathbf{b}}_{t+h} = \mathbb{E}\left[\mathbf{B}_{t+h} \mid \mathcal{I}_{t,b}, \widehat{\mathbf{u}}_{t+h}\right] = \widehat{\mathbf{b}}_{t+h} + \mathbf{G}(\widehat{\mathbf{u}}_{t+h} - \mathbf{A}\widehat{\mathbf{b}}_{t+h}) \tag{11}$$

$$\mathrm{Var}\left[\mathbf{B}_{t+h} \mid \mathcal{I}_{t,b}, \widehat{\mathbf{u}}_{t+h}\right] = k_h\left(\widehat{\boldsymbol{\Sigma}}_{B,1} - \mathbf{G}(\mathbf{A}\widehat{\boldsymbol{\Sigma}}_{B,1} \ + \mathbf{M}_1^T)\right) \tag{12}$$

Thus the adjustment applied to the base forecasts is proportional to $(\widehat{\mathbf{u}}_{t+h} - \mathbf{A}\widehat{\mathbf{b}}_{t+h})$, i.e. the difference between the prior mean and the uncertain observation (i.e., the forecasts) of the upper time series. The term $(\widehat{\mathbf{u}}_{t+h} - \mathbf{A}\widehat{\mathbf{b}}_{t+h})$ is called the *incoherence* of the base forecasts in [18]. The mean of the reconciled bottom time series does not depend on $k_h$, while the variance does.

The reconciled point forecast and the covariance for the entire hierarchy are:

$$\mathbb{E}[\boldsymbol{Y}_{t+h} \mid \mathcal{I}_{t,b}, \widehat{\mathbf{u}}_{t+h}] = \widetilde{\mathbf{y}}_{t+h} = \mathbf{S}\widetilde{\mathbf{b}}_{t+h} \tag{13}$$

$$\mathrm{Var}\left[\boldsymbol{Y}_{t+h} \mid \mathcal{I}_{t,b}, \widehat{\mathbf{u}}_{t+h}\right] = \mathbf{S}\,\mathrm{Var}\left[\mathbf{B}_{t+h} \mid \mathcal{I}_{t,b}, \widehat{\mathbf{u}}_{t+h}\right]\mathbf{S}^T. \tag{14}$$

### 3.2 Related works and optimality of the reconciliation

Bayes' rule is a well-known tool for information fusion [5, Sec. 2], and we apply it for the first time for forecast reconciliation. We will later prove that the posterior mean of our approach yields the same point predictions of MinT.

Yet, our algorithm additionally provides the predictive distribution for the entire hierarchy; we thus call it pMinT, where p stands for *probabilistic*. We also contribute a novel reconciliation approach based on the linear-Gaussian (LG) model, which is obtained by setting $\mathbf{M} = \mathbf{0}$ in the definition of $\mathbf{G}$. Both pMinT and LG are thus probabilistic reconciliation algorithms.

We point out for the first time a link between the reconciliation problem and the Kalman filter, whose state-update equation can be derived from the linear-Gaussian model [14]. In particular, Equation (11) has the same structure of the state-update of a Kalman filter. According to the definition of $\mathbf{G}$ of Eq.(10), the LG reconciliation corresponds to the standard Kalman filter [16, Chap. 5], while the pMinT reconciliation corresponds to a generalized Kalman filter which assumes correlation between the noise of the state and the noise of the output [16, Chap. 7.1]. Thus future research could explore the literature of the Kalman filter in order to borrow ideas for the reconciliation problem.

The optimality of our approach can be informally proven by considering that it yields the posterior mean of the reconciled forecasts, which is the minimizer of the quadratic loss under the Gaussian assumption [12, Chap. 5.7]. Moreover, it has

the same equation of the state-update step of the Kalman filter, which provably minimizes the mean squared error of the estimates without any distributional assumption [16, Chap. 5]. In Sec. (4.1) we will moreover show that our point predictions correspond to those of MinT, which have been proven to be the minimizer of the mean-squared error.

### 3.3 The covariance matrices $\widehat{\boldsymbol{\Sigma}}_{B,1}$ and $\widehat{\boldsymbol{\Sigma}}_{U,1}$

The element $(i,j)$ of $\widehat{\boldsymbol{\Sigma}}_{B,1}$ is the covariance $\mathrm{Cov}(B^i_{t+1}, B^j_{t+1} \mid \mathcal{I}_{t,b}) = \mathrm{Cov}(B^i_{t+1}, B^j_{t+1} \mid \mathbf{B}_{1:t} = \mathbf{b}_{1:t})$, where $\mathbf{B}_{1:t} = \mathbf{b}_{1:t}$ denotes a realization $\mathbf{b}_{1:t}$ of $\mathbf{B}_{1:t}$. Yet we only have one observation of $B^i_{t+1}, B^j_{t+1}$ *conditional* on $\mathcal{I}_{t,b}$, which prevents estimating the covariance. We can overcome the problem with the following result, which shows that we can approximate $\widehat{\boldsymbol{\Sigma}}_{B,1}$ by computing the covariance of the residuals.

Let us consider the vectors of bottom time series $\mathbf{B}_1, \dots, \mathbf{B}_t$ and the conditional expectation $\widehat{\mathbf{B}}_{t+1} = \mathbb{E}[\mathbf{B}_{t+1} \mid \mathbf{B}_1, \dots, \mathbf{B}_t] = \mathbb{E}[\mathbf{B}_{t+1} \mid \mathbf{B}_{1:t}]$. Note that $\widehat{\mathbf{B}}_{t+1}$ is a random vector as we have not yet observed $\mathbf{B}_i$, $i = 1, \dots, t$. In this section we show

$$\mathbb{E}[\mathrm{Cov}(B^i_{t+1}, B^j_{t+1} \mid \mathbf{B}_{1:t})] = \mathrm{Cov}(E^i_{t+1}, E^j_{t+1}) \qquad i,j = 1, \dots, n, \tag{15}$$

where $E^i_{t+1} := B^i_{t+1} - \widehat{B}^i_{t+1}$, $i = 1, \dots, n$ denotes the residual of the model fitted on the $i$-th time series, for the forecast horizon $t+1$. If we observe $\mathcal{I}_{t,b}$, then we can approximate $\mathrm{Cov}(B^i_{t+1}, B^j_{t+1} \mid \mathcal{I}_{t,b})$ with $\mathrm{Cov}(B^i_{t+1} - \widehat{b}^i_{t+1}, B^j_{t+1} - \widehat{b}^j_{t+1})$, the covariance of the residuals of the models fitted on the bottom time series.

Consider now the conditional covariance on the left side of eq. (15), we have

$$\begin{aligned}
\mathrm{Cov}(B^i_{t+1}, B^j_{t+1} \mid \mathbf{B}_{1:t}) &= \mathbb{E}\left[ \left(B^i_{t+1} - \mathbb{E}[B^i_{t+1} \mid \mathbf{B}_{1:t}]\right) \left(B^j_{t+1} - \mathbb{E}[B^j_{t+1} \mid \mathbf{B}_{1:t}]\right) \mid \mathbf{B}_{1:t} \right] \\
&= \mathbb{E}\left[ \left(B^i_{t+1} - \widehat{B}^i_{t+1}\right) \left(B^j_{t+1} - \widehat{B}^j_{t+1}\right) \mid \mathbf{B}_{1:t} \right]
\end{aligned}$$

By taking the expectation on both sides we obtain

$$\begin{aligned}
\mathbb{E}[\mathrm{Cov}(B^i_{t+1}, B^j_{t+1} \mid \mathbf{B}_{1:t})] &= \mathbb{E}\left[ \mathbb{E}\left[ \left(B^i_{t+1} - \widehat{B}^i_{t+1}\right) \left(B^j_{t+1} - \widehat{B}^j_{t+1}\right) \mid \mathbf{B}_{1:t} \right] \right] \\
&= \mathbb{E}\left[ \left(B^i_{t+1} - \widehat{B}^i_{t+1}\right) \left(B^j_{t+1} - \widehat{B}^j_{t+1}\right) \right] \\
&= \mathrm{Cov}\left(B^i_{t+1} - \widehat{B}^i_{t+1}, B^j_{t+1} - \widehat{B}^j_{t+1}\right) = \mathrm{Cov}\left(E^i_{t+1}, E^j_{t+1}\right)
\end{aligned}$$

We thus estimate the covariance matrix $\widehat{\boldsymbol{\Sigma}}_{B,1}$ using the covariance of the residuals of the models fitted on the bottom time series.

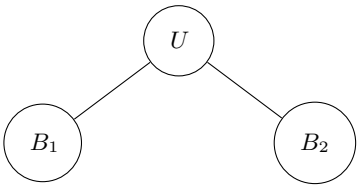
**Computation of $\widehat{\boldsymbol{\Sigma}}_{U,1}$** According to Eq. (7),

$$\boldsymbol{\varepsilon}^u_{t+1} = \mathbf{A}\mathbf{B}_{t+1} - \widehat{\mathbf{U}}_{t+1},$$

whose variances and covariances can be readily computed from the residuals.

## 4 Reconciliation of a simple hierarchy

We now illustrate how the base forecasts interact during the reconciliation of a simple hierarchy. We consider a hierarchy constituted by two bottom time series ($B_1$ and $B_2$) and an upper time series $U$.




The base forecast for the bottom time series are the point forecasts $\widehat{b}_1$ and $\widehat{b}_2$ with variances $\sigma_1^2$ and $\sigma_2^2$. The prior beliefs about $B_1$ and $B_2$ are:

$$\begin{pmatrix} B_1 \\ B_2 \end{pmatrix} \sim N\left[\begin{pmatrix} \widehat{b}_1 \\ \widehat{b}_2 \end{pmatrix}, k_h \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{1,2} & \sigma_2^2, \end{pmatrix}\right]$$

where for simplicity we remove the forecast horizon $(t+h)$ from the notation.

The summing matrix is:

$$\mathbf{S} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{A} \\ 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

We start considering the simpler case of reconciliation via the LG algorithm. The matrix $\mathbf{G}$ is:

$$\mathbf{G} = \widehat{\boldsymbol{\Sigma}}_{B,1}\mathbf{A}^T(\widehat{\boldsymbol{\Sigma}}_{U,1} + \mathbf{A}\widehat{\boldsymbol{\Sigma}}_{B,1}\mathbf{A}^T)^{-1} = \frac{1}{\sigma_u^2 + \sigma_1^2 + \sigma_2^2 + 2\sigma_{1,2}} \begin{bmatrix} \sigma_1^2 + \sigma_{1,2} \\ \sigma_2^2 + \sigma_{1,2}, \end{bmatrix}$$

since $\mathbf{A} = [1\ 1]$, $\widehat{\boldsymbol{\Sigma}}_{U,1} = \sigma_u^2$ and moreover:

$$\begin{aligned} &\widehat{\boldsymbol{\Sigma}}_{B,1}\mathbf{A}^T = [\sigma_1^2 + \sigma_{1,2} \ \ \sigma_{1,2} + \sigma_2^2]^T, \\ &\mathbf{A}^T\widehat{\boldsymbol{\Sigma}}_{B,1}\mathbf{A} = \sigma_1^2 + \sigma_2^2 + 2\sigma_{1,2}, \\ &\widehat{\boldsymbol{\Sigma}}_{U,1} + \mathbf{A}\widehat{\boldsymbol{\Sigma}}_{B,1}\mathbf{A}^T = \sigma_u^2 + \sigma_1^2 + \sigma_2^2 + 2\sigma_{1,2}. \end{aligned}$$

Note that $\mathbf{G}$ does not depend on $h$, as also shown in Eq. (10).

The reconciled bottom forecasts are:

$$\widetilde{\mathbf{b}} = \widehat{\mathbf{b}} + \mathbf{G}(\widehat{u} - \mathbf{A}\widehat{\mathbf{b}}), \tag{16}$$

where $\widehat{u}$ is the base forecast for $U$.

The reconciled bottom forecast can be written as:

$$\widetilde{\mathbf{b}} = \begin{bmatrix} \widehat{b}_1 \\ \widehat{b}_2 \end{bmatrix} + \begin{bmatrix} \sigma_1^2 + \sigma_{1,2} \\ \sigma_2^2 + \sigma_{1,2} \end{bmatrix} \frac{\widehat{u} - \mathbf{A}^T \widehat{\mathbf{b}}}{\sigma_u^2 + \sigma_1^2 + \sigma_2^2 + 2\sigma_{1,2}} \tag{17}$$

Eq. (17) shows that the adjustment applied to the base forecasts depends on $\sigma_u^2$. If $\sigma_u^2$ is large the adjustment is small, since the upper forecast is not informative. If on the contrary $\sigma_u^2 = 0$, the sum of the reconciled bottom forecasts is forced to match $\hat{u}$, i.e., $\tilde{b}_1 + \tilde{b}_2 = \hat{u}$ (this can be shown by re-working Eq. (17)).

We now show that the reconciled bottom forecast are a linear combination of the base forecasts. Let us define:

$$g_1 = \frac{\sigma_1^2 + \sigma_{1,2}}{\sigma_u^2 + \sigma_1^2 + \sigma_2^2 + 2\sigma_{1,2}} \quad g_2 = \frac{\sigma_2^2 + \sigma_{1,2}}{\sigma_u^2 + \sigma_1^2 + \sigma_2^2 + 2\sigma_{1,2}} \tag{18}$$

After some algebra we obtain:

$$\widetilde{\mathbf{b}} = \begin{bmatrix} \widehat{b}_1 \left(1 - \frac{\sigma_1^2+\sigma_{1,2}}{\sigma_u^2+\sigma_1^2+\sigma_2^2+2\sigma_{1,2}}\right) + (\widehat{u} - \widehat{b}_2) \frac{\sigma_1^2+\sigma_{1,2}}{\sigma_u^2+\sigma_1^2+\sigma_2^2+2\sigma_{1,2}} \\ \widehat{b}_2 \left(1 - \frac{\sigma_2^2+\sigma_{1,2}}{\sigma_u^2+\sigma_1^2+\sigma_2^2+2\sigma_{1,2}}\right) + (\widehat{u} - \widehat{b}_1) \frac{\sigma_2^2+\sigma_{1,2}}{\sigma_u^2+\sigma_1^2+\sigma_2^2+2\sigma_{1,2}} \end{bmatrix} = \begin{bmatrix} \widehat{b}_1 (1 - g_1) + (\widehat{u} - \widehat{b}_2) g_1 \\ \widehat{b}_2 (1 - g_2) + (\widehat{u} - \widehat{b}_1) g_2, \end{bmatrix} \tag{19}$$

Thus $\widetilde{b}_1$ is a weighted average of two estimates: $\widehat{b}_1$ and $(\widehat{u} - \widehat{b}_2)$; the weight of $\widehat{b}_1$ decreases with $\sigma_1^2$ and increases with $(\sigma_2^2 + \sigma_u^2)$.

The reconciliation carried out by pMinT is similar to what already discussed, once we adopt $g_1^*$ and $g_2^*$ in place of $g_1$ and $g_2$:

$$g_1^* = \frac{\sigma_1^2 + \sigma_{1,2} - \sigma_{u,1}}{\sigma_u^2 + \sigma_1^2 + \sigma_2^2 + 2\sigma_{1,2} - 2\sigma_{u,1} - 2\sigma_{u,2}} \tag{20}$$

$$g_2^* = \frac{\sigma_2^2 + \sigma_{1,2} - \sigma_{u,2}}{\sigma_u^2 + \sigma_1^2 + \sigma_2^2 + 2\sigma_{1,2} - 2\sigma_{u,1} - 2\sigma_{u,2}} \tag{21}$$

Thus pMinT accounts also for the cross-covariances $\sigma_{u,1}, \sigma_{u,2}$ between the bottom time series and of noise affecting the forecasts for the upper time series.

### 4.1 Relationship with MinT

Our reconciled bottom time series can be written as:

$$\begin{aligned} \widetilde{\mathbf{b}} &= \widehat{\mathbf{b}} + \mathbf{G}(\widehat{\mathbf{u}} - \mathbf{A}\widehat{\mathbf{b}}) = (\mathbf{I} - \mathbf{G}\mathbf{A})\widehat{\mathbf{b}} + \mathbf{G}\widehat{\mathbf{u}} \\ &= \left[\mathbf{G}\ (\mathbf{I} - \mathbf{G}\mathbf{A})\right] \begin{bmatrix} \widehat{\mathbf{u}} \\ \widehat{\mathbf{b}} \end{bmatrix} = \mathbf{P}_h \widehat{\mathbf{y}}. \end{aligned} \tag{22}$$

The matrix $\mathbf{P}_h$ of pMinT is thus:

$$\mathbf{P}_{pMinT,h} = \left[\mathbf{G}\ (\mathbf{I} - \mathbf{G}\mathbf{A})\right] \tag{23}$$

**Proposition 1.** *The matrices $\mathbf{P}_h$ of MinT and pMinT are equivalent. The proof is given in the supplementary material.*

## 5 Experiments

In this paper we take a probabilistic point of view; we thus assess the reconciled predictive distributions rather than the point forecasts. Our metric is the *energy score* (ES), a scoring rule for multivariate distributions [7]. The ES is the multivariate generalization of the continuous ranked probability score (CRPS), which is obtained by integrating the Brier score over the predictive distribution of the forecast [7]. Let $\mathbf{y}$ be the actual multivariate observation, and let us assume that we have $k$ samples $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$, from the multivariate predictive distribution $F$. The energy score is:

$$\mathrm{ES}(\mathbf{y}, F) = \frac{1}{k}\sum_{i=1}^{k} \|\mathbf{x}_i - \mathbf{y}\| - \frac{1}{2k^2}\sum_{i=1}^{k}\sum_{j=1}^{k} \|\mathbf{x}_i - \mathbf{x}_j\| \tag{24}$$

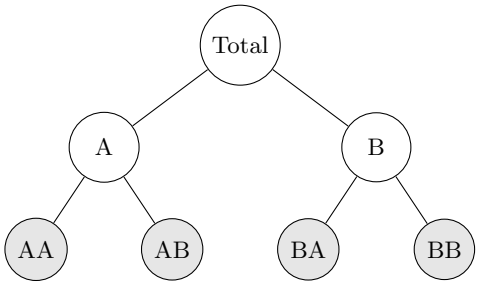
The energy score is a loss function: the lower, the better. We consider three methods for probabilistic reconciliation: probabilistic bottom-up (BU), LG and pMinT. We did not find any package implementing the algorithms of [17, 13]; thus we did not include them in our comparison. We estimate all the covariance matrices via the shrinkage estimator [15].

*Base forecasts* We consider two time series models to compute the base forecasts. The first is *ets*, which fits different exponential smoothing variants and eventually performs model selection via AICc. The second method is auto.arima. It first decides how to differentiate the time series to make it stationary; then, it looks for the best arma model which fits the stationary time series, performing model selection via AICc. Both approaches performed well in forecasting competitions; they are available from the *forecast* package [10] for R. In all simulations, we compute forecasts up to $h$=4. As no particular pattern exists in the relative performance of the methods as $h$ varies, we present the results averaged over $h$=1,2,3,4.

*Setting $k_h$* We are unaware of previous studies on how to set $k_h$. In this paper we compare two heuristics, acknowledging that this remains an open problem. The two options are $k_h$=$h$ and $k_h$=1. The choice $k_h$=$h$ is based on the following approximation. The variance of $\widehat{y}_{t+1}$ around $y_{t+1}$ is $\sigma^2$; assuming the independence of the errors, the variance of $\widehat{y}_{t+1}$ around $y_{t+h}$ is $h\sigma^2$. The approximation lies in the fact that we are modeling the variance of $\widehat{y}_{t+1}$ (not $\widehat{y}_{t+h}$) around $y_{t+h}$.

Instead, the option $k_h = 1$ keeps the variance fixed with $h$. This represents the short-term behavior of models which contain only seasonal terms and no autoregressive terms. For instance, when dealing with a monthly time series, the variance of such models is constant up to $h$=12.

*Code* The code of our experiments is available at: `https://github.com/iamthejao/BayesianReconciliation`.




### 5.1 Synthetic data

We generate synthetic data sets using the hierarchy above, previously considered in the experiments of [18]. We simulate the four bottom time series as AR(1) processes, drawing their parameters uniformly from the stationary region. The noises of the bottom time series at each time instant are correlated, multivariate Gaussian distributed, with mean $\mu = [0, 0, 0, 0]^T$ and covariance:

$$\Sigma = \begin{bmatrix} 5 & 3 & 2 & 1 \\ 3 & 5 & 2 & 1 \\ 2 & 2 & 5 & 3 \\ 1 & 1 & 3 & 5 \end{bmatrix}.$$

Thus $\Sigma$ enforces a stronger correlation between time series which have the same parents. At each time instant $t$ we add the noise $\eta_t \sim N(0, 10)$ to the time series AA and BA and the noise $(-\eta_t)$ to the time series AB and BB. In this way we simulate noisy bottom time series ($\eta_t$ and $-\eta_t$ cancel out when dealing with the upper time series) which can be encountered in real cases when several disaggregations are applied to the total time series. We consider the following length $T$ of the time series: $\{50; 100; 1000\}$. For each value of $T$ we perform 1000 simulations. The averaged energy scores are given in Tab. 1; in each cell we report the lower energy score between the case $k_h$=$h$ and $k_h$=1.

Since the time series are stationary, they basically fluctuate around their mean. In this case the magnitude of the incoherence is generally limited, allowing also the bottom-up reconciliation to be competitive. The ES of both pMinT and LG is on average 1.5% smaller than that of BU. We also note an advantage of LG over pMinT for small $T$, and instead the reverse for large $T$; this might be the effect of the additional covariances estimated by pMinT (see $\sigma_{u,1}$ and $\sigma_{u,2}$ in Sec.4).

### 5.2 Experiments with real data sets

We consider two hierarchical time series: *infantgts* and *tourism*. Both are *grouped* time series, which is a generalization of hierarchical time series. In particular the time series of a given level are always sums of some bottom time series, but they are not necessarily sums of time series of the adjacent lower level.

Table 1: Mean energy score, averaged over 1000 simulations and over $h$=1,2,3,4. In each row we highlight the lower result.

| $T$ | method | BU | pMinT | LG |
|---|---|---|---|---|
| 50 | arima | 9.7 | 9.5 | **9.4** |
| 50 | ets | 9.9 | 9.8 | **9.7** |
| 100 | arima | 9.1 | **9.0** | **9.0** |
| 100 | ets | 9.5 | 9.4 | **9.3** |
| 1000 | arima | 8.8 | **8.7** | 8.8 |
| 1000 | ets | 9.5 | **9.3** | 9.4 |

*Infantgts* The *infantgts* is available within the **hts** [8] package for R. It contains infant mortality counts in Australia, disaggregated by sex and by eight different states. Each time series contains 71 yearly observations, covering the period 1933-2003. The bottom level contains 16 time series (8 states x 2 genders). The second level contains 2 time series: the counts of males and females, aggregated over the states. The third level sums males and females in each state, yielding 8 time series (one for each state). The fourth level is the total.

*Tourism* The *tourism* data set regards the number of nights spent by Australians away from home. It is available in raw format from `https://robjhyndman.com/publications/MinT/`. The time series cover the period 1998–2016 with monthly frequency. There are 304 bottom time series, referring to 76 regions and 4 purposes. The first level sums over the purposes, yielding 76 time series (one for each region); such values are further aggregated into macro-zones (27 time series) and states (7 time series). Other levels of the hierarchy aggregate the bottom time series of the same zone (yielding 108 time series: 27 zones x 4 purposes), which are then further aggregated into 28 time series (7 states x 4 purposes) and then 4 time series (4 purposes). The last level is the total. Overall the hierarchy contains 555 time series.

We repeat 50 times the following procedure: split the time series into training and test, using a different split point; compute the base forecasts up to $h$=4; reconcile the forecasts. The reconciliation is independently computed for each $h$. Each value of Tab. 2 is thus the average over 200 experiments (50 training/test splits $\times$ $h$=1,2,3,4). On *infantgts*, all the three reconciliation methods perform better with $k_h = h$, probably because the variance of the fitted time series models steadily increases with $h$. On the contrary, on *tourism* all reconciliation algorithms perform better with $k_h$=1; in this case most models only contain the seasonal part. The rows referring to the best values of $k_h$ are highlighted in Tab. 2.

We call *setup* the combination of a data set and a forecasting method, such as <infangts,arima >. The pMinT algorithm yields the lowest energy score in most

| dset | $k_h$ | method | BU | pMinT | LG |
|---|---|---|---|---|---|
| infantgts | 1 | arima | **334.1** | 346.9 | 348.5 |
| | 1 | ets | 334.0 | **320.0** | 334.7 |
| | *h* | arima | **327.2** | 335.1 | 331.0 |
| | *h* | ets | 328.2 | **313.7** | 318.7 |
| tourism | 1 | arima | 2,737.6 | **2,412.0** | 2,547.4 |
| | 1 | ets | 2,496.0 | **2,403.7** | 2,520.1 |
| | *h* | arima | 2,785.3 | **2,380.3** | 2,448.2 |
| | *h* | ets | 2,527.1 | **2,353.6** | 2,410.3 |

Table 2: Averaged energy scores. Each cell is the average over 200 reconciliations (50 different training/test splits × $h$=1,2,3,4) . The rows corresponding to the best-performing values of $k_h$ are highlighted.

setups; in the next section we check whether the differences between methods are significant.

*Statistical analysis* For each setup we perform a significance tests for each pair of algorithms (pMinT vs BU, LG vs BU and pMinT vs LG), using the Bayesian signed-rank test [2], which returns the posterior probability of a method having lower median energy score than another (Tab. 3). Such posterior probabilities are numerically equivalent to (1 - *p-value*), where *p-value* is the p-value of the one-sided frequentist signed-rank test.

In most setups (Tab. 3) high posterior probabilities (implying low p-values) support the hypothesis of the pMinT having lower energy score than both BU and LG; moreover they also support the hypothesis of LG having lower energy score than BU.

| | | | *Posterior probabilities* | | |
|---|---|---|---|---|---|
| dset | $k_h$ | method | pMinT <BU | pMinT <LG | LG <BU |
| infantgts | 1 | arima | 0.24 | 0.02 | 0.47 |
| | | ets | 1.00 | 0.85 | 1 |
| | h | arima | 0.89 | 0.00 | 1.00 |
| | | ets | 1.00 | 0.00 | 1 |
| tourism | 1 | arima | 1.00 | 1.00 | 1.00 |
| | | ets | 1.00 | 1.00 | 0.25 |
| | h | arima | 1.00 | 1.00 | 1.00 |
| | | ets | 1.00 | 1.00 | 1.00 |

Table 3: Posterior probability of the Bayesian signed rank test.

*Meta-analysis* We now perform a meta-analysis for for each pair of algorithms across the different setups, adopting the Poisson-binomial approach [11, 4]. Consider for instance pMinT vs BU. We model each setup as a Bernoulli trial, whose possible outcomes are the victory of pMinT or BU. The probability of pMinT winning is taken for each setup from Tab. 3 (the probability of BU winning is just its complement to 1). We then repeat 10,000 simulations, in which we draw the outcome of each setup according to the probabilities of Tab. 3.

We now report the probability of each method outperforming another method in more than half the setups, based on out of 10,000 simulations. Both pMinT and LG wins in more than half the setup with probability 1 against BU. Moreover, there is 0.85 probability of pMinT winning in more than half of the setups against LG. We thus recommend pMinT as a general default method for probabilistic reconciliation.

## 6 Conclusions

We have derived two algorithms (pMinT and LG) based on Bayes' rule for probabilistic reconciliation. We have also shown a didactic example which clarifies how base forecast and their variances interact during the reconciliation, In general pMinT yields better predictive distributions and thus we recommend it as a default. The LG method can be anyway an interesting alternative when dealing with small sample sizes. Future research could borrow ideas from the extensive literature of the Kalman filter, based on the link we pointed out between reconciliation and Kalman filter

## Acknowledgements

We acknowledge support from grant n. 407540_167199 / 1 from Swiss NSF (NRP 75 Big Data).

## References

1. Athanasopoulos, G., Gamakumara, P., Panagiotelis, A., Hyndman, R., M., A.: Hierarchical forecasting. In: Macroeconomic Forecasting in the Era of Big Data. pp. 689–719 (2020)
2. Benavoli, A., Corani, G., Demšar, J., Zaffalon, M.: Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis. The Journal of Machine Learning Research **18**(1), 2653–2688 (2017)
3. Bishop, C.: Pattern Recognition and Machine Learning. Springer (2006)
4. Corani, G., Benavoli, A., Demšar, J., Mangili, F., Zaffalon, M.: Statistical comparison of classifiers through Bayesian hierarchical modelling. Machine Learning **106**(11), 1817–1837 (2017)
5. Durrant-Whyte, H., Henderson, T.C.: Multisensor data fusion. Springer Handbook of Robotics pp. 585–610 (2008)

6. Gamakumara, P., Panagiotelis, A., Athanasopoulos, G., Hyndman, R.: Probabilisitic forecasts in hierarchical time series. Tech. rep., Monash University, Department of Econometrics and Business Statistics (2018)

7. Gneiting, T., Raftery, A.E.: Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association **102**(477), 359–378 (2007)

8. Hyndman, R., Lee, A., Wang, E., Wickramasuriya, S.: hts: Hierarchical and Grouped Time Series (2018), `https://CRAN.R-project.org/package=hts`, R package version 5.1.5

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

10. Hyndman, R.J., Khandakar, Y.: Automatic time series forecasting: the forecast package for R. Journal of Statistical Software **3**(27), 1–22 (2008)

11. Lacoste, A., Laviolette, F., Marchand, M.: Bayesian comparison of machine learning algorithms on single and multiple datasets. In: Proc. 15th Int. Conf. on Artificial Intelligence and Statistics. pp. 665–675 (2012)

12. Murphy, K.P.: Machine learning: a probabilistic perspective. MIT press (2012)

13. Park, M., Nassar, M.: Variational Bayesian inference for forecasting hierarchical time series. In: ICML 2014 Workshop on Divergence Methods for Probabilistic Inference. pp. 1–6 (2014)

14. Roweis, S., Ghahramani, Z.: A unifying review of linear Gaussian models. Neural computation **11**(2), 305–345 (1999)

15. Schäfer, J., Strimmer, K.: A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical applications in genetics and molecular biology **4**(1) (2005)

16. Simon, D.: Optimal state estimation: Kalman, H infinity and nonlinear approaches. John Wiley & Sons (2006)

17. Taieb, S.B., Taylor, J.W., Hyndman, R.J.: Coherent probabilistic forecasts for hierarchical time series. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. vol. 70, pp. 3348–3357 (2017)

18. Wickramasuriya, S.L., Athanasopoulos, G., Hyndman, R.J.: Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. Journal of the American Statistical Association **114**(526), 804–819 (2019)

# A Supplementary material

*Proof (Proposition 1: equivalence between the point forecasts of minT and pMinT).*
We denote the matrix $\mathbf{P}$ of minT as $\mathbf{P}_{MinT}$. According to [18, Theorem 1], $\mathbf{P}_{MinT}$ is given by:

$$\mathbf{P}_{MinT} = \begin{bmatrix} \mathbf{0}_{n\times(m-n)} \\ \mathbf{I}_n \end{bmatrix} - \begin{bmatrix} \mathbf{0}_{n\times(m-n)} \\ \mathbf{I}_n \end{bmatrix} \mathbf{W}_h \begin{bmatrix} \mathbf{I}_{(m-n)} \\ -\mathbf{A}^T \end{bmatrix} \left( \begin{bmatrix} \mathbf{I}_{(m-n)} \\ -\mathbf{A}^T \end{bmatrix}^T \mathbf{W}_h \begin{bmatrix} \mathbf{I}_{(m-n)} \\ -\mathbf{A}^T \end{bmatrix} \right)^{-1} \begin{bmatrix} \mathbf{I}_{(m-n)} \\ -\mathbf{A}^T \end{bmatrix}^T \tag{25}$$

The covariance matrix $\mathbf{W}_h$ has the following structure:

$$\begin{aligned} \mathbf{W}_h &= \mathbb{E}[(\mathbf{y}_{t+h} - \widehat{\mathbf{y}}_{t+h})(\mathbf{y}_{t+h} - \widehat{\mathbf{y}}_{t+h})^T \mid \mathcal{I}_t] \\ &= \mathbb{E}[\begin{pmatrix} \mathbf{u}_{t+h} - \widehat{\mathbf{u}}_{t+h} \\ \mathbf{b}_{t+h} - \widehat{\mathbf{b}}_{t+h} \end{pmatrix} \begin{pmatrix} \mathbf{u}_{t+h} - \widehat{\mathbf{u}}_{t+h} \\ \mathbf{b}_{t+h} - \widehat{\mathbf{b}}_{t+h} \end{pmatrix}^T \mid \mathcal{I}_t] \\ &= \begin{bmatrix} k_h \widehat{\boldsymbol{\Sigma}}_{U,1} & -k_h \mathbf{M}^T \\ -k_h \mathbf{M} & k_h \widehat{\boldsymbol{\Sigma}}_{B,1} \end{bmatrix}. \end{aligned}$$

Note that the cross-covariances in the matrix $\mathbf{W}_h$ have a negative sign in front of $\mathbf{M}$ and $\mathbf{M}^T$. This is because we compute $\mathbf{M} = \mathrm{Cov}(\mathbf{B}_{t+h}, \boldsymbol{\varepsilon}^u_{t+h} \mid \mathcal{I}_{t,b}) = \mathbb{E}[(\mathbf{B}_{t+h} - \widehat{\mathbf{b}}_{t+h})(\boldsymbol{\varepsilon}^u_{t+h})^T \mid \mathcal{I}_{t,b}]$ while the cross covariance in $\mathbf{W}_h$ is $\mathbb{E}[(\mathbf{b}_{t+h} - \widehat{\mathbf{b}}_{t+h})(\mathbf{u}_{t+h} - \widehat{\mathbf{u}}_{t+h})^T \mid \mathcal{I}_{t,b}] = \mathbb{E}[(\mathbf{b}_{t+h} - \widehat{\mathbf{b}}_{t+h})(-\boldsymbol{\varepsilon}_{t+h})^T \mid \mathcal{I}_{t,b}]$.

The inner parenthesis in Eq. (25), then results in

$$\begin{aligned} \begin{bmatrix} \mathbf{I}_{(m-n)} \\ -\mathbf{A}^T \end{bmatrix}^T \begin{bmatrix} k_h \widehat{\boldsymbol{\Sigma}}_{U,1} & -k_h \mathbf{M}^T \\ -k_h \mathbf{M} & k_h \widehat{\boldsymbol{\Sigma}}_{B,1} \end{bmatrix} \begin{bmatrix} \mathbf{I}_{(m-n)} \\ -\mathbf{A}^T \end{bmatrix} &= \begin{bmatrix} k_h(\widehat{\boldsymbol{\Sigma}}_{U,1} + \mathbf{AM}) & -k_h(\mathbf{M}^T - \mathbf{A}\widehat{\boldsymbol{\Sigma}}_{B,1}) \end{bmatrix} \begin{bmatrix} \mathbf{I}_{(m-n)} \\ -\mathbf{A}^T \end{bmatrix} \\ &= k_h(\widehat{\boldsymbol{\Sigma}}_{U,1} + \mathbf{AM} + \mathbf{M}^T\mathbf{A}^T + \mathbf{A}\widehat{\boldsymbol{\Sigma}}_{B,1}\mathbf{A}^T) \end{aligned}$$

Moreover we have

$$\begin{aligned} \mathbf{P}_{MinT} &= \begin{bmatrix} \mathbf{0}_{n\times(m-n)} \\ \mathbf{I}_n \end{bmatrix} - k_h \begin{bmatrix} -\mathbf{M} & \widehat{\boldsymbol{\Sigma}}_{B,1} \end{bmatrix} \begin{bmatrix} \mathbf{I}_{(m-n)} \\ -\mathbf{A}^T \end{bmatrix} k_h^{-1} \left( \widehat{\boldsymbol{\Sigma}}_{U,1} + \mathbf{AM} + \mathbf{M}^T\mathbf{A}^T + \mathbf{A}\widehat{\boldsymbol{\Sigma}}_{B,1}\mathbf{A}^T \right)^{-1} \begin{bmatrix} \mathbf{I}_{(m-n)} \\ -\mathbf{A}^T \end{bmatrix}^T \\ &= \begin{bmatrix} \mathbf{0}_{n\times(m-n)} \\ \mathbf{I}_n \end{bmatrix} + (\mathbf{M} + \widehat{\boldsymbol{\Sigma}}_{B,1}\mathbf{A}^T) \left( \widehat{\boldsymbol{\Sigma}}_{U,1} + \mathbf{AM} + \mathbf{M}^T\mathbf{A}^T + \mathbf{A}\widehat{\boldsymbol{\Sigma}}_{B,1}\mathbf{A}^T \right)^{-1} \begin{bmatrix} \mathbf{I}_{(m-n)} \\ -\mathbf{A}^T \end{bmatrix}^T \\ &= \begin{bmatrix} \mathbf{0}_{n\times(m-n)} \\ \mathbf{I}_n \end{bmatrix} + \begin{bmatrix} \mathbf{G} \\ -\mathbf{GA} \end{bmatrix} = \begin{bmatrix} \mathbf{G} \\ \mathbf{I}_n - \mathbf{GA}, \end{bmatrix} \end{aligned}$$

which corresponds to the $\mathbf{P}$ matrix we derived for pMinT in Eq. (23).