Chapter 1
Game-theoretically Optimal Reconciliation of Contemporaneous Hierarchical Time Series Forecasts

Tim van Erven and Jairo Cugliari

Abstract In hierarchical time series (HTS) forecasting, the hierarchical relation between multiple time series is exploited to make better forecasts. This hierarchical relation implies one or more aggregate consistency constraints that the series are known to satisfy. Many existing approaches, like for example bottom-up or top-down forecasting, therefore attempt to achieve this goal in a way that guarantees that the forecasts will also be aggregate consistent. We propose to split the problem of HTS into two independent steps: first one comes up with the best possible forecasts for the time series without worrying about aggregate consistency; and then a reconciliation procedure is used to make the forecasts aggregate consistent. We introduce a Game-Theoretically OPtimal (GTOP) reconciliation method, which is guaranteed to only improve any given set of forecasts. This opens up new possibilities for constructing the forecasts. For example, it is not necessary to assume that bottom-level forecasts are unbiased, and aggregate forecasts may be constructed by regressing both on bottom-level forecasts and on other covariates that may only be available at the aggregate level. We illustrate the benefits of our approach both on simulated data and on real electricity consumption data.

1.1 Introduction

The general setting of *hierarchical time series* (HTS) forecasting has been extensively studied because of its applications to, among others, inventory management for companies [Fliedner, 1999], euro-area macroeconomic studies [Lütkepohl, 2009], forecasting Australian domestic tourism [Hyndman et al., 2011], and balanc-

Tim van Erven
Université Paris-Sud, 91405 Orsay Cedex, France, and INRIA e-mail: tim@timvanerven.nl

Jairo Cugliari
Laboratoire ERIC, Université Lumière Lyon2, 69676 Bron Cedex, France, e-mail: Jairo.Cugliari@univ-lyon2.fr

ing the national budget of states [Stone et al., 1942, Byron, 1978]. As a consequence of the recent deployment of smart grids and autodispatchable sources, HTS have also been introduced in electricity demand forecasting [Borges et al., 2013], which is essential for electricity companies to reduce electricity production cost and take advantage of market opportunities.

A Motivating Example: Electricity Forecasting The electrical grid induces a hierarchy in which customer demand is viewed at increasing levels of aggregation. One may organize this hierarchy in different ways, but in any case the demand of individual customers is at the bottom, and the top level represents the total demand for the whole system. Depending on the modelling purpose, intermediate levels of aggregation may represent groups of clients that are tied together by geographical proximity, tariff contracts, similar consumption structure or other criteria.

Whereas demand data were previously available only for the whole system, they are now also available at regional (intermediate) levels or even at the individual level, which makes it possible to forecast electricity demand at different levels of aggregation. To this end, it is not only necessary to extend existing prediction models to lower levels of the customer hierarchy, but also to deal with the new possibilities and constraints that are introduced by the hierarchical organization of the predictions. In particular, it may be required that the sum of lower-level forecasts is equal to the prediction for the whole system. This was demanded, for example, in the Global Energy Forecasting Competition 2012 [Hong et al., 2013], and it also makes intuitive sense that the forecasts should sum in the same way as the real data. Moreover, we show in Theorems 1 and 2 below that this requirement, if enforced using a general method that we will introduce, can only improve the forecasts.

Hierarchical Time Series Electricity demand data that are organized in a customer hierarchy, are a special case of what is known in the literature as *contemporaneous* HTS: each node in the hierarchy corresponds to a time series, and, at any given time, the value of a time series higher up is equal to the sum of its constituent time series. In contrast, there also exist *temporal* HTS, in which time series are aggregated over periods of time, but we will not consider those in this work. For both types of HTS, the question of whether it is better to predict an aggregate time series directly or to derive forecasts from predictions for its constituent series has received a lot of attention, although the consensus appears to be that there is no clear-cut answer. (See [Fliedner, 1999, Lütkepohl, 2009] for surveys.) A significant theoretical effort has also been made to understand the probability structure of contemporaneous HTS when the constituent series are *auto-regressive moving average* (ARMA) models [Granger, 1988].

HTS Forecasting The most common methods used for hierarchical time series forecasting are *bottom-up*, *top-down* and *middle-out* [Fliedner, 1999, Borges et al., 2013]. The first of these concentrates on the prediction of all the components and uses the sum of these predictions as a forecast of the whole. The second one predicts the top level aggregate and then splits up the prediction into the components according to proportions that may be estimated, for instance, from historical proportions

in the time series. The middle out strategy is a combination of the first two: one first obtains predictions at some level of the hierarchy; then one uses the bottom-up strategy to forecast the upper levels and top-down to forecast the lower levels.

As observed by Hyndman et al. [2011], all three methods can be viewed as linear mappings from a set of initial forecasts for the time series to *reconciled* estimates that are *aggregate consistent*, which means that the sum of the forecasts of the components of an hierarchical time series is equal to the forecast of the whole. A more sophisticated linear mapping may be obtained by setting up a linear regression problem in which the initial forecasts are viewed as noisy observations of the expected values of the time series [Byron, 1978] (see Section 1.2.3). In this approach, which goes back to Stone, Champernowne, and Meade [1942], it is then inescapable to assume that the initial forecasts are unbiased estimates, so that the noise has mean zero. Assuming furthermore that the covariance matrix Σ of the noise can be accurately estimated for each time step, the outcomes for the time series can be estimated using a *generalized least-squares* (GLS) method, which solves the linear regression problem with aggregate consistency constraints on the solution.

Although the assumption of unbiased initial forecasts rules out using any type of regularized estimator (like, for instance, the LASSO [Tibshirani, 1996] which we consider in Section 1.3.1), it might still be manageable in practice. The difficulty with GLS, however, is estimating Σ, which might be possible on accounting data by laboriously tracing back all the sources of variance in the estimates [Chen, 2006], but does not seem feasible in our motivating example of electricity demand forecasting. (Standard estimators like those of White [1980] or MacKinnon and White [1985] do not apply, because they estimate an average of Σ over time instead of its value at the current time step.) Alternatively, it has therefore been proposed to make an additional assumption about the covariances of the initial forecasts that allows estimation of Σ to be sidestepped [Hyndman et al., 2011], but it is not clear when we can expect this assumption to hold (see Section 1.2.3).

Our Contribution Considering the practical difficulties in applying GLS, and the limited modelling power of bottom-up, top-down, middle-out methods, we try to approach HTS forecasting in a slightly different way. All these previous approaches have been restricted by combining the requirement of aggregate consistency with the goal of sharing information between hierarchical levels. Instead, we propose to separate these steps, which leads to an easier way of thinking about the problem. As our main contribution, we will introduce a Game-Theoretically OPtimal (GTOP) reconciliation method to map any given set of forecasts, which need not be aggregate consistent, to new aggregate consistent forecasts that are guaranteed to be at least as good. As the GTOP method requires no assumptions about the probability structure of the time series or the nature of the given set of forecasts, it leaves a forecaster completely free to use the prediction method of their choice at all levels of the hierarchy without worrying about aggregate consistency or theoretical restrictions like unbiasedness of their forecasts. As illustrated in Section 1.3.2, taking aggregate consistency out of the equation allows one to go beyond simple bottom-up, top-down or middle-out estimators, and consider estimators that use more complicated

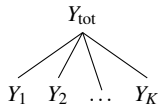

Fig. 1.1 A two-level hierarchical time series structure

regression structures, in the same spirit as those considered by Lütkepohl [2009, Section 5.3].

Outline In the next section, we present the GTOP method and formally relate it to the GLS approach. Then, in Section 1.3, we demonstrate how GTOP may be applied with forecasts that do not satisfy the traditional unbiasedness assumption, first on simulated data, and then on real electricity demand data. Finally, Section 1.4 provides an extensive discussion.

1.2 Game-theoretically Optimal Reconciliation

We will now introduce the GTOP method, which takes as input a set of forecasts, which need not be aggregate consistent, and produces as output new aggregate consistent forecasts that are guaranteed to be at least as good. In Section 1.2.1, we first present the method for the simplest possible hierarchies, which are composed of two levels only, and then, in Section 1.2.2, we explain how the procedure generalizes in a straightforward way to arbitrary hierarchies. Proofs and computational details are postponed until the end of Section 1.2.2. Finally, in Section 1.2.3, we show how GTOP reconciliation may formally be interpreted as a special case of GLS, although the quantities involved have different interpretations.

1.2.1 Two-level Hierarchies

For two-level hierarchies, we will refer to the lower levels as *regions*, in reference to our motivating application of electricity demand forecasting, even though for other applications the lower levels might correspond to something else. Suppose there are K such regions, and we are not only interested in forecasting the values of a time series $(Y_k[t])_{t=1,2,\dots}$ for each individual region $k = 1,\dots,K$, but also in forecasting the sum of the regions $(Y_{\text{tot}}[t])_{t=1,2,\dots}$, where

$$Y_{\text{tot}}[t] = \sum_{k=1}^{K} Y_k[t] \qquad \text{for all } t, \tag{1.1}$$

as illustrated by Figure 1.1.

Having observed the time series for times $1,\ldots,t$, together with possible independent variables, we will be concerned with making predictions for their values at time $\tau > t$, but to avoid clutter, we will drop the time index $[\tau]$ from our notation whenever it is sufficiently clear from context. Thus, for any region k, let $\hat{Y}_k \equiv \hat{Y}_k[\tau]$ be the prediction for $Y_k \equiv Y_k[\tau]$, and let $\hat{Y}_{\text{tot}} \equiv \hat{Y}_{\text{tot}}[\tau]$ be the prediction for $Y_{\text{tot}} \equiv Y_{\text{tot}}[\tau]$. Then we evaluate the quality of our prediction for region k by the *squared loss*

$$\ell_k(Y_k, \hat{Y}_k) = a_k(Y_k - \hat{Y}_k)^2,$$

where $a_k > 0$ is a weighting factor that is determined by the operational costs associated with prediction errors in region k. (We give some guidelines for the choice of these weighting factors in Section 1.4.1.) Similarly, our loss in predicting the sum of the regions is

$$\ell_{\text{tot}}(Y_{\text{tot}}, \hat{Y}_{\text{tot}}) = a_{\text{tot}}(Y_{\text{tot}} - \hat{Y}_{\text{tot}})^2,$$

with $a_{\text{tot}} > 0$. Let $\mathbf{Y} = (Y_1, \ldots, Y_K, Y_{\text{tot}})$ and $\hat{\mathbf{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_K, \hat{Y}_{\text{tot}})$. Then, all together, our loss at time τ is

$$\ell(\mathbf{Y}, \hat{\mathbf{Y}}) = \sum_{k=1}^{K} \ell_k(Y_k, \hat{Y}_k) + \ell_{\text{tot}}(Y_{\text{tot}}, \hat{Y}_{\text{tot}}).$$

Aggregate Inconsistency In predicting the total Y_{tot}, we might be able to take advantage of covariates that are only available at the aggregate level or there might be noise that cancels out between regions, so that we have to anticipate that $\hat{Y}_{\text{tot}}$ may be a better prediction of Y_{tot} than simply the sum of the regional predictions $\sum_{k=1}^{K} \hat{Y}_k$, and generally we may have $\hat{Y}_{\text{tot}} \neq \sum_{k=1}^{K} \hat{Y}_k$.[1] In light of (1.1), allowing such an aggregate inconsistency between the regional predictions and the prediction for the total would intuitively seem suboptimal. More importantly, for operational reasons it is sometimes not even allowed. For example, in the Global Energy Forecasting Competition 2012 [Hong et al., 2013], it was required that the sum of the regional predictions $\hat{Y}_1, \ldots, \hat{Y}_K$ were always equal to the prediction for the total $\hat{Y}_{\text{tot}}$. Or, if the time series represent next year's budgets for different departments, then the budget for the whole organization must typically be equal to the sum of the budgets for the departments.

We are therefore faced with a choice between two options. The first is that we might try to adjust our prediction methods to avoid aggregate inconsistency. But this would introduce complicated dependencies between our prediction methods for the different regions and for the total, and as a consequence it might make our predictions worse. So, alternatively, we might opt to remedy the problem in a post-processing step: first we come up with the best possible predictions $\hat{\mathbf{Y}}$ without worrying about any potential aggregate inconsistency, and then we map these predictions to new predictions $\widetilde{\mathbf{Y}} = (\widetilde{Y}_1, \ldots, \widetilde{Y}_K, \widetilde{Y}_{\text{tot}})$, which are *aggregate consistent*:

[1] It has also been suggested that the *central limit theorem* (CLT) implies that Y_{tot} should be more smooth than the individual regions Y_k [Borges et al., 2013], and might therefore be easier to predict.

$$\widetilde{Y}_{\text{tot}} = \sum_{k=1}^{K} \widetilde{Y}_k.$$

This is the route we will take in this paper. In fact, it turns out that, for the right mapping, the loss of $\widetilde{\mathbf{Y}}$ will always be smaller than the loss of $\hat{\mathbf{Y}}$, no matter what the actual data $\mathbf{Y}$ turn out to be, which provides a formal justification for the intuition that aggregate inconsistent predictions should be avoided.

Mapping to Aggregate Consistent Predictions To map any given predictions $\hat{\mathbf{Y}}$ to aggregate consistent predictions $\widetilde{\mathbf{Y}}$, we will use a game-theoretic set-up that is reminiscent of the game-theoretic approach to online learning [Cesa-Bianchi and Lugosi, 2006]. In this formulation, we will choose our predictions $\widetilde{\mathbf{Y}}$ to achieve the minimum in the following minimax optimization problem:

$$V = \min_{\widetilde{\mathbf{Y}} \in \mathscr{A}} \max_{\mathbf{Y} \in \mathscr{A} \cap \mathscr{B}} \Big\{ \ell(\mathbf{Y}, \widetilde{\mathbf{Y}}) - \ell(\mathbf{Y}, \hat{\mathbf{Y}}) \Big\}. \tag{1.2}$$

(The sets $\mathscr{A}$ and $\mathscr{B}$ will be defined below.) This may be interpreted as the *Game-Theoretically OPtimal* (GTOP) move in a zero-sum game in which we first choose $\widetilde{\mathbf{Y}}$, then the data $\mathbf{Y}$ are chosen by an adversary, and finally the pay-off is measured by the difference in loss between $\widetilde{\mathbf{Y}}$ and the given predictions $\hat{\mathbf{Y}}$. The result is that we will choose $\widetilde{\mathbf{Y}}$ to guarantee that $\ell(\mathbf{Y}, \widetilde{\mathbf{Y}}) - \ell(\mathbf{Y}, \hat{\mathbf{Y}})$ is at most V *no matter what the data* $\mathbf{Y}$ *are*. Satisfyingly, we shall see below that $V \leq 0$, so that the new predictions $\widetilde{\mathbf{Y}}$ are always at least as good as the original predictions $\hat{\mathbf{Y}}$.

We have left open the definitions of the sets $\mathscr{A}$ and $\mathscr{B}$, which represent the domains for our predictions and the data. The former of these will represent the set of vectors that are aggregate consistent:

$$\mathscr{A} = \Big\{ (X_1, \ldots, X_K, X_{\text{tot}}) \in \mathbb{R}^{K+1} \mid X_{\text{tot}} = \sum_{k=1}^{K} X_k \Big\}.$$

By definition, both our predictions $\widetilde{\mathbf{Y}}$ and the data $\mathbf{Y}$ must be aggregate consistent, so they are restricted to lie in $\mathscr{A}$. In addition, we introduce the set $\mathscr{B}$, which allows us to specify any other information we might have about the data. In the simplest case, we may let $\mathscr{B} = \mathbb{R}^{K+1}$ so that $\mathscr{B}$ imposes no constraints, but if, for example, prediction intervals $[\hat{Y}_k - B_k, \hat{Y}_k + B_k]$ are available for the given predictions, then we may take advantage of that knowledge and define

$$\mathscr{B} = \Big\{ (X_1, \ldots, X_K, X_{\text{tot}}) \in \mathbb{R}^{K+1} \mid X_k \in [\hat{Y}_k - B_k, \hat{Y}_k + B_k] \text{ for } k = 1, \ldots, K \Big\}. \tag{1.3}$$

We could also add a prediction interval for $\hat{Y}_{\text{tot}}$ as long as we take care that all our prediction intervals together do not contradict aggregate consistency of the data. In general, we will require that $\mathscr{B} \subseteq \mathbb{R}^{K+1}$ is a *closed* and *convex* set, and $\mathscr{A} \cap \mathscr{B}$ must be non-empty so that $\mathscr{B}$ does not contradict aggregate consistency.

GTOP Predictions as a Projection Let $\|\mathbf{X}\| = (\sum_{i=1}^{d} X_i^2)^{1/2}$ denote the L2-norm of a vector $\mathbf{X} \in \mathbb{R}^d$ for any dimension d. Then the total loss may succinctly be written as

$$\ell(\mathbf{Y}, \hat{\mathbf{Y}}) = \|A\mathbf{Y} - A\hat{\mathbf{Y}}\|^2, \tag{1.4}$$

where $A = \mathrm{diag}(\sqrt{a_1}, \ldots, \sqrt{a_K}, \sqrt{a_{\mathrm{tot}}})$ is a diagonal $(K+1) \times (K+1)$ matrix that accounts for the weighting factors. In view of the loss, it is quite natural that the GTOP predictions turn out to be equal to the L2-projection

$$\widetilde{\mathbf{Y}}_{\mathrm{proj}} = \operatorname*{arg\,min}_{\widetilde{\mathbf{Y}} \in \mathscr{A} \cap \mathscr{B}} \|A\hat{\mathbf{Y}} - A\widetilde{\mathbf{Y}}\|^2 \tag{1.5}$$

of $\hat{\mathbf{Y}}$ unto $\mathscr{A} \cap \mathscr{B}$ after scaling all dimensions according to A.

Theorem 1 (GTOP: Two-level Hierarchies). *Suppose that $\mathscr{B}$ is a closed, convex set and that $\mathscr{A} \cap \mathscr{B}$ is not empty. Then the projection $\widetilde{\mathbf{Y}}_{\mathrm{proj}}$ uniquely exists, the value of* (1.2) *is*

$$V = -\|A\widetilde{\mathbf{Y}}_{\mathrm{proj}} - A\hat{\mathbf{Y}}\|^2 \leq 0,$$

and the GTOP predictions are $\widetilde{\mathbf{Y}} = \widetilde{\mathbf{Y}}_{\mathrm{proj}}$.

Thus, in a metric that depends on the loss, GTOP makes the minimal possible adjustment of the given predictions $\hat{\mathbf{Y}}$ to make them consistent with what we know about the data. Moreover, the fact that $V \leq 0$ implies that the GTOP predictions are at least as good as the given predictions:

$$\ell(\mathbf{Y}, \widetilde{\mathbf{Y}}_{\mathrm{proj}}) \leq \ell(\mathbf{Y}, \hat{\mathbf{Y}}) \qquad \text{for any data } \mathbf{Y} \in \mathscr{A} \cap \mathscr{B}.$$

Theorem 1 will be proved as a special case of Theorem 2 in the next section.

Example 1. If $\mathscr{B} = \mathbb{R}^{K+1}$ does not impose any constraints, then the GTOP predictions are

$$\begin{aligned}
\widetilde{Y}_{\mathrm{proj},k} &= \hat{Y}_k \ + \frac{\frac{1}{a_k}}{\sum_{i=1}^{K} \frac{1}{a_i} + \frac{1}{a_{\mathrm{tot}}}} \Delta \qquad \text{for } k = 1, \ldots, K, \\
\widetilde{Y}_{\mathrm{proj,tot}} &= \hat{Y}_{\mathrm{tot}} - \frac{\frac{1}{a_{\mathrm{tot}}}}{\sum_{i=1}^{K} \frac{1}{a_i} + \frac{1}{a_{\mathrm{tot}}}} \Delta,
\end{aligned}$$

where $\Delta = \hat{Y}_{\mathrm{tot}} - \sum_{k=1}^{K} \hat{Y}_k$ measures by how much $\hat{\mathbf{Y}}$ violates aggregate consistency. In particular, if the given predictions $\hat{\mathbf{Y}}$ are already aggregate consistent, i.e. $\hat{Y}_{\mathrm{tot}} = \sum_{k=1}^{K} \hat{Y}_k$, then the GTOP predictions are the same as the given predictions: $\widetilde{\mathbf{Y}}_{\mathrm{proj}} = \hat{\mathbf{Y}}$.

Example 2. If $\mathscr{B}$ consists of the prediction intervals specified in (1.3), then the extreme values $B_1 = \ldots = B_K = 0$ make the GTOP predictions exactly equal to those of the bottom-up forecaster.

Example 3. If $\mathscr{B}$ defines prediction intervals as in (1.3) and $a_1 = \cdots = a_K = a$ and $B_1 = \cdots = B_K = B$, then the GTOP predictions are

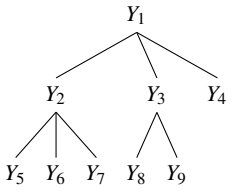

Fig. 1.2 Example of a multi-level hierarchical time series structure

$$\widetilde{Y}_{\text{proj},k} = \hat{Y}_k + \Big[\frac{\frac{1}{a}}{\frac{K}{a}+\frac{1}{a_{\text{tot}}}}\Delta\Big]_B \quad \text{for } k = 1,\dots,K,$$
$$\widetilde{Y}_{\text{proj,tot}} = \sum_{k=1}^{K} \widetilde{Y}_{\text{proj},k},$$

where $[x]_B = \max\big\{-B, \min\{B,x\}\big\}$ denotes clipping x to the interval $[-B,B]$ and $\Delta = \hat{Y}_{\text{tot}} - \sum_{k=1}^{K} \hat{Y}_k$.

In general the GTOP predictions $\widetilde{\mathbf{Y}}_{\text{proj}}$ do not have a closed-form solution, but, as long as $\mathscr{B}$ can be described by a finite set of inequality constraints, they can be computed using quadratic programming. The details will be discussed at the end of the next section, which generalizes the two-level hierarchies introduced so far to arbitrary summation constraints.

1.2.2 General Summation Constraints

One might view (1.1) as forecasting $K+1$ time series, which are ordered in a hierarchy with two levels, in which the time series $(Y_1[t])$, …, $(Y_K[t])$ for the regions are at the bottom, and their total $(Y_{\text{tot}}[t])$ is at the top (see Figure 1.1). More generally, one might imagine having a multi-level hierarchy of any finite number of time series $(Y_1[t]),\dots,(Y_M[t])$, which are organised in a tree T that represents the hierarchy of aggregation consistency requirements. For example, in Figure 1.2 the time series $(Y_1[t])$ might be the expenditure of an entire organisation, the time series $(Y_2[t]),(Y_3[t])$, and $(Y_4[t])$ might be the expenditures in different subdivisions within the organization, time series $(Y_5[t]),(Y_6[t])$ and $(Y_7[t])$ might represent the expenditures in departments within subdivision $(Y_2[t])$, and similarly $(Y_8[t])$ and $(Y_9[t])$ would be the expenditures in departments within $(Y_3[t])$.

The discussion from the previous section directly extends to multi-level hierarchies as follows. For each time series $m = 1,\dots,M$, let $c(m) \subset \{1,\dots,M\}$ denote the set of its children in T. Then aggregate consistency generalizes to the constraint

$$\mathscr{A} = \Big\{(X_1,\dots,X_M) \in \mathbb{R}^M \mid X_m = \sum_{i\in c(m)} X_i \text{ for all } m \text{ such that } c(m) \text{ is non-empty}\Big\}.$$

Remark 1. We note that all the constraints $X_m = \sum_{i \in c(m)} X_i$ in $\mathscr{A}$ are *linear equality constraints*. In fact, in all the subsequent developments, including Theorem 2, we can allow $\mathscr{A}$ to be *any set* of linear equality constraints, as long as they are internally consistent, so that $\mathscr{A}$ is not empty. In particular, we could even allow two (or more) predictions for the same time series by regarding the first prediction as a prediction for a time series $(Y_m[t])$ and the second as a prediction for a separate time series $(Y_{m'}[t])$ with the constraint that $Y_m[t] = Y_{m'}[t]$. To keep the exposition focussed, however, we will not explore these possibilities in this paper.

Having defined the structure of the hierarchical time series through $\mathscr{A}$, any additional information we may have about the data can again be represented by choosing a convex, closed set $\mathscr{B} \subseteq \mathbb{R}^M$ which is such that $\mathscr{A} \cap \mathscr{B}$ is non-empty. In particular, $\mathscr{B} = \mathbb{R}^M$ represents having no further information, and prediction intervals can be represented analogously to (1.3) if they are available.

As in the two-level hierarchy, let $\hat{\mathbf{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_M)$ be the original (potentially aggregate inconsistent) predictions for the time series $\mathbf{Y} = (Y_1, \ldots, Y_M)$ at a given time τ. We assign weighting factors $a_m > 0$ to each of the time series $m = 1, \ldots, M$, and we redefine the diagonal matrix $A = \mathrm{diag}(\sqrt{a_1}, \ldots, \sqrt{a_M})$, so that we may write the total loss as in (1.4). Then the GTOP predictions $\widetilde{\mathbf{Y}} = (\widetilde{Y}_1, \ldots, \widetilde{Y}_M)$ are still defined as those achieving the minimum in (1.2), and the L2-projection $\widetilde{\mathbf{Y}}_{\mathrm{proj}}$ is as defined in (1.5).

Theorem 2 (GTOP: Multi-level Hierarchies). *The exact statement of Theorem 1 still holds for the more general definitions for multi-level hierarchies in this section.*

The proof of Theorems 1 and 2 fundamentally rests on the Pythagorean inequality, which is illustrated by Figure 1.3. In fact, this inequality is not restricted to the squared loss we use in this paper, but holds for any loss that is based on a Bregman divergence [Cesa-Bianchi and Lugosi, 2006, Section 11.2], so the proof would go through in exactly the same way for such other losses. For example, the Kullback-Leibler divergence, which measures the difference between two probability distributions, is also a Bregman divergence.

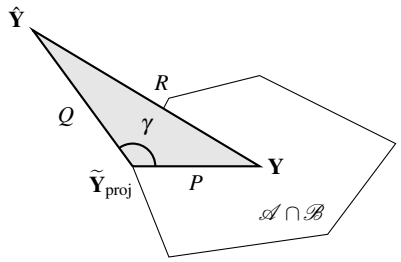

Fig. 1.3 Illustration of the Pythagorean inequality $P^2 + Q^2 \leq R^2$, where $P = \|A\mathbf{Y} - A\widetilde{\mathbf{Y}}_{\mathrm{proj}}\|$, $Q = \|A\widetilde{\mathbf{Y}}_{\mathrm{proj}} - A\hat{\mathbf{Y}}\|$ and $R = \|A\mathbf{Y} - A\hat{\mathbf{Y}}\|$. Convexity of $\mathscr{A} \cap \mathscr{B}$ ensures that $\gamma \geq 90^\circ$.

Lemma 1 (Pythagorean Inequality). *Suppose that $\mathscr{B}$ is a closed, convex set and that $\mathscr{A} \cap \mathscr{B}$ is non-empty. Then the projection $\widetilde{\mathbf{Y}}_{\text{proj}}$ exists and is unique, and*

$$\|A\mathbf{Y} - A\widetilde{\mathbf{Y}}_{\text{proj}}\|^2 + \|A\widetilde{\mathbf{Y}}_{\text{proj}} - A\hat{\mathbf{Y}}\|^2 \leq \|A\mathbf{Y} - A\hat{\mathbf{Y}}\|^2 \qquad \textit{for all } \mathbf{Y} \in \mathscr{A} \cap \mathscr{B}.$$

Proof. The lemma is an instance of the generalized Pythagorean inequality [Cesa-Bianchi and Lugosi, 2006, Section 11.2] for the Bregman divergence corresponding to the Legendre function $F(\mathbf{X}) = \|A\mathbf{X}\|^2$, which is strictly convex (as required) because all entries of the matrix A are strictly positive. (The set $\mathscr{A}$ is a hyperplane, so it is closed and convex by construction. The assumptions of the lemma therefore ensure that $\mathscr{A} \cap \mathscr{B}$ is closed, convex and non-empty.) □

Proof (Theorem 2). Let $f(\mathbf{Y}, \widetilde{\mathbf{Y}}) = \ell(\mathbf{Y}, \widetilde{\mathbf{Y}}) - \ell(\mathbf{Y}, \hat{\mathbf{Y}})$. We will show that $(\widetilde{\mathbf{Y}}_{\text{proj}}, \widetilde{\mathbf{Y}}_{\text{proj}})$ is a saddle-point for f, which implies that playing $\widetilde{\mathbf{Y}}_{\text{proj}}$ is the optimal strategy for both players in the zero-sum game and that

$$V = \min_{\widetilde{\mathbf{Y}} \in \mathscr{A}} \max_{\mathbf{Y} \in \mathscr{A} \cap \mathscr{B}} f(\mathbf{Y}, \widetilde{\mathbf{Y}}) = \max_{\mathbf{Y} \in \mathscr{A} \cap \mathscr{B}} \min_{\widetilde{\mathbf{Y}} \in \mathscr{A}} f(\mathbf{Y}, \widetilde{\mathbf{Y}}) = f(\widetilde{\mathbf{Y}}_{\text{proj}}, \widetilde{\mathbf{Y}}_{\text{proj}}) = -\|A\widetilde{\mathbf{Y}}_{\text{proj}} - A\hat{\mathbf{Y}}\|^2$$

[Rockafellar, 1970, Lemma 36.2], which is to be shown.

To prove that $(\widetilde{\mathbf{Y}}_{\text{proj}}, \widetilde{\mathbf{Y}}_{\text{proj}})$ is a saddle-point, we need to show that neither player can improve their pay-off by changing their move. To this end, we first observe that, by the Pythagorean inequality (Lemma 1),

$$f(\mathbf{Y}, \widetilde{\mathbf{Y}}_{\text{proj}}) = \|A\mathbf{Y} - A\widetilde{\mathbf{Y}}_{\text{proj}}\|^2 - \|A\mathbf{Y} - A\hat{\mathbf{Y}}\|^2 \leq -\|A\widetilde{\mathbf{Y}}_{\text{proj}} - A\hat{\mathbf{Y}}\|^2 = f(\widetilde{\mathbf{Y}}_{\text{proj}}, \widetilde{\mathbf{Y}}_{\text{proj}})$$

for all $\mathbf{Y} \in \mathscr{B} \cap \mathscr{A}$. It follows that the maximum is achieved by $\mathbf{Y} = \widetilde{\mathbf{Y}}_{\text{proj}}$. Next, we also have

$$\arg\min_{\widetilde{\mathbf{Y}} \in \mathscr{A}} f(\widetilde{\mathbf{Y}}_{\text{proj}}, \widetilde{\mathbf{Y}}) = \arg\min_{\widetilde{\mathbf{Y}} \in \mathscr{A}} \|A\widetilde{\mathbf{Y}}_{\text{proj}} - A\widetilde{\mathbf{Y}}\|^2 = \widetilde{\mathbf{Y}}_{\text{proj}},$$

which completes the proof. □

Efficient Computation For special cases, like the examples in the previous section, the GTOP projection $\widetilde{\mathbf{Y}}_{\text{proj}}$ sometimes has a closed form. In general, no closed-form solution may be available, but $\widetilde{\mathbf{Y}}_{\text{proj}}$ can still be computed by finding the solution to the quadratic program

$$\begin{aligned} \min_{\widetilde{\mathbf{Y}}} \quad & \|A\hat{\mathbf{Y}} - A\widetilde{\mathbf{Y}}\|^2 \\ \text{subject to} \quad & \widetilde{\mathbf{Y}} \in \mathscr{A} \cap \mathscr{B}. \end{aligned}$$

Since $\mathscr{A}$ imposes only equality constraints, this quadratic program can be solved efficiently as long as the further constraints imposed by $\mathscr{B}$ are manageable. In particular, if $\mathscr{B}$ imposes only linear inequality constraints, like, for example, in (1.3), then the solution can be found efficiently using interior point methods [Lobo et al., 1998] or using any of the alternatives suggested by Hazan et al. [2007, Section 4].

The experiments in Section 1.3 were all implemented using the `quadprog` package for the `R` programming language, which turned out to be fast enough.

1.2.3 Formal Relation to Generalized Least-squares

As discussed in the introduction, HTS has been modelled as a problem of linear regression in the economics literature [Byron, 1978]. It is interesting to compare this approach to GTOP, because the two turn out to be very similar, except that the quantities involved have different interpretations. The linear regression approach models the predictions as functions of the means of the real data

$$\hat{\mathbf{Y}}[\tau] = \mathbb{E}\left\{\mathbf{Y}[\tau]\right\} + \varepsilon[\tau]$$

that are perturbed by a noise vector $\varepsilon[\tau] = (\varepsilon_1[\tau], \ldots, \varepsilon_M[\tau])$, where all distributions and expectations are conditional on all previously observed values of the time series. Then it is assumed that the predictions are *unbiased estimates*, so that the noise variables all have mean zero, and the true means $\mathbb{E}\{\mathbf{Y}[\tau]\}$ can be estimated using the *generalized least-squares* (GLS) estimate

$$\begin{aligned} \min_{\widetilde{\mathbf{Y}}} \quad & (\hat{\mathbf{Y}} - \widetilde{\mathbf{Y}})^\top \Sigma^{-1} (\hat{\mathbf{Y}} - \widetilde{\mathbf{Y}}) \\ \text{subject to} \quad & \widetilde{\mathbf{Y}} \in \mathscr{A}, \end{aligned} \tag{1.6}$$

where $\Sigma \equiv \Sigma[\tau]$ is the $M \times M$ covariance matrix for the noise $\varepsilon[\tau]$ [Byron, 1978]. This reveals an interesting superficial relation between the GTOP forecasts and the GLS estimates: if

$$\Sigma^{-1} = A^\top A \qquad \text{and} \qquad \mathscr{B} = \mathbb{R}^M, \tag{1.7}$$

then the two coincide! However, the interpretation of A and Σ^{-1} is completely different, and the two procedures serve different purposes: whereas GLS tries to address both reconciliation and the goal of sharing information between hierarchical levels at the same time, the GTOP method is only intended to do reconciliation and requires a separate procedure to share information. The case where the two methods coincide is therefore only a formal coincidence, and one should not assume that the choice $\Sigma^{-1} = A^\top A$ will adequately take care of sharing information between hierarchical levels!

Ordinary Least-squares Given the difficulty of estimating Σ, Hyndman et al. [2011] propose an assumption that allows them to sidestep estimation of Σ altogether: they show that, under their assumption, the GLS estimate reduces to the *Ordinary Least-squares* (OLS) estimate obtained from (1.6) by the choice

$$\Sigma = I,$$

where I is the identity matrix. Via (1.7) it then follows that the OLS and GTOP forecasts formally coincide when we take all the weighting factors in the definition of the loss to be equal: $a_1 = \ldots = a_M$, and let $\mathscr{B} = \mathbb{R}^M$. Consequently, for two-level hierarchies, OLS can be computed as in Example 1.

The assumption proposed by Hyndman et al. [2011] is that, at time τ, the covariance $\mathrm{Cov}(\hat{Y}_m, \hat{Y}_{m'})$ of the predictions for any two time series decomposes as

$$\mathrm{Cov}(\hat{Y}_m, \hat{Y}_{m'}) = \sum_{\substack{i \in S(m) \\ j \in S(m')}} \mathrm{Cov}(\hat{Y}_i, \hat{Y}_j) \quad \text{for all } m, m', \tag{1.8}$$

where $S(m) \subseteq \{1, \ldots, M\}$ denotes the set of bottom-level time series out of which Y_m is composed. That is, $Y_m = \sum_{i \in S(m)} Y_i$ with Y_i childless (i.e. $c(i) = \emptyset$) for all $i \in S(m)$.

Although the OLS approach appears to work well in practice (see Section 1.3.2), it is not obvious when we can expect (1.8) to hold. Hyndman et al. [2011] motivate it by pointing out that (1.8) would hold exactly if the forecasts would be exactly aggregate consistent (i.e. $\hat{\mathbf{Y}} \in \mathscr{A}$). Since it is reasonable to assume that the forecasts will be approximately aggregate consistent, it then also seems plausible that (1.8) will hold approximately. However, this motivation seems insufficient, because reasoning *as if* the forecasts are aggregate consistent leads to conclusions that are too strong: if $\hat{\mathbf{Y}} \in \mathscr{A}$, then any instance of GLS would give the same answer, so it would not matter which Σ we used, and in the experiments in Section 1.3 we see that this clearly does matter.

We therefore prefer to view OLS rather as a special case of GTOP, which will work well when all the weighting factors in the loss are equal and the constraints in $\mathscr{B}$ are vacuous.

1.3 Experiments

As discussed above, the GTOP method only solves the reconciliation part of HTS forecasting; it does not prescribe how to construct the original predictions $\hat{\mathbf{Y}}$. We will now illustrate how GTOP might be used in practice, taking advantage of the fact that it does not require the original predictions $\hat{\mathbf{Y}}$ to be unbiased. First, in Section 1.3.1, we present a toy example with simulated data, which nevertheless illustrates many of the difficulties one might encounter on real data. Then, in Section 1.3.2, we apply GTOP to real electricity demand data, which motivated its development.

1.3.1 Simulation Study

We use GTOP with prediction intervals as in (1.3). We will compare to bottom-up forecasting, and also to the OLS method described in Section 1.2.3, because it appears to work well in practice (see Section 1.3.2) and it is one of the few methods

available that does not require estimating any parameters. We do not compare to top-down forecasting, because estimating proportions in top-down forecasting is troublesome in the presence of independent variables (see Section 1.4.2).

Data We consider a two-level hierarchy with two regions, and simulate data according to

$$Y_1[t] = \beta_{1,0} + \beta_{1,1} X[t] + \varepsilon_1[t] \qquad Y_2[t] = \beta_{2,0} + \beta_{2,1} X[t] + \varepsilon_2[t]$$

where $(X[t])$ is an independent variable, $\beta_1 = (\beta_{1,0}, \beta_{1,1})$ and $\beta_2 = (\beta_{2,0}, \beta_{2,1})$ are coefficients to be estimated, and $(\varepsilon_1[t])$ and $(\varepsilon_2[t])$ are noise variables. We will take $\beta_1 = \beta_2 = (1,5)$, and let

$$\varepsilon_1[t] = \tau \vartheta_1[t] + \sigma \upsilon[t] \qquad \varepsilon_2[t] = \tau \vartheta_2[t] - \sigma \upsilon[t] \qquad \text{for all } t,$$

where $\vartheta_1[t]$, $\vartheta_2[t]$ and $\upsilon[t]$ are uniformly distributed on $[-1,1]$, independently over t and independently of each other, and τ and σ are scale parameters, for which we will consider different values. Notice that the noise that depends on $\upsilon[t]$ cancels from the total $Y_{\text{tot}}[t] = Y_1[t] + Y_2[t]$, which makes the total easier to predict than the individual regions. We sample a train set of size 100 for the fixed design $(X[t])_{t=1,\ldots,100} = (1/100, 2/100, \ldots, 1)$ and a test set of the same size for $(X[t])_{t=101,\ldots,200} = (1 + 1/100, \ldots, 2)$.

Fitting Models on the Train Set Based on the train set, we find estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ of the coefficients β_1 and β_2 by applying the LASSO [Tibshirani, 1996] separately for each of the two regions, using cross-validation to calibrate the amount of penalization. Then we predict $Y_1[\tau]$ and $Y_2[\tau]$ by

$$\hat{Y}_1[\tau] = \hat{\beta}_{1,0} + \hat{\beta}_{1,1} X[\tau] \qquad \hat{Y}_2[\tau] = \hat{\beta}_{2,0} + \hat{\beta}_{2,1} X[\tau].$$

Remark 2. In general, it is not guaranteed that forecasting the total $Y_{\text{tot}}[\tau]$ directly will give better predictions than the bottom-up forecast [Lütkepohl, 2009]. Consequently, if the bottom-up forecast is the best we can come up with, then that is how we should define our prediction for the total, and no further reconciliation is necessary!

If we would use the LASSO directly to predict the total $Y_{\text{tot}}[\tau]$, then, in light of Remark 2, it might not do better than simply using the *bottom-up* forecast $\hat{Y}_1[\tau] + \hat{Y}_2[\tau]$. We can be sure to do better than the bottom-up forecaster, however, by *adding our regional forecasts* $\hat{Y}_1[\tau]$ *and* $\hat{Y}_2[\tau]$ *as covariates*, such that we fit $Y_{\text{tot}}[\tau]$ by

$$\beta_{\text{tot},0} + \beta_{\text{tot},1} X[\tau] + \beta_{\text{tot},2} \hat{Y}_1[\tau] + \beta_{\text{tot},3} \hat{Y}_2[\tau], \tag{1.9}$$

where $\beta_{\text{tot}} = (\beta_{\text{tot},0}, \beta_{\text{tot},1}, \beta_{\text{tot},2}, \beta_{\text{tot},3})$ are coefficients to be estimated. For $\beta_{\text{tot}} = (0,0,1,1)$ this would exactly give the bottom-up forecast, but now we can also obtain different estimates if the data tell us to use different coefficients. However, to be conservative and take advantage of the prior knowledge that the bottom-up forecast is often quite good, we introduce prior knowledge into the LASSO by regularizing

by

$$|\beta_{\text{tot},0}| + |\beta_{\text{tot},1}| + |\beta_{\text{tot},2} - 1| + |\beta_{\text{tot},3} - 1| \tag{1.10}$$

instead of its standard regularization by $|\beta_{\text{tot},0}| + |\beta_{\text{tot},1}| + |\beta_{\text{tot},2}| + |\beta_{\text{tot},3}|$, which gives it a preference for coefficients that are close to those of the bottom-up forecast. Thus, from the train set, we obtain estimates $\hat{\beta}_{\text{tot}} = (\hat{\beta}_{\text{tot},0}, \hat{\beta}_{\text{tot},1}, \hat{\beta}_{\text{tot},2}, \hat{\beta}_{\text{tot},3})$ for β_{tot}, and we predict $Y_{\text{tot}}[\tau]$ by

$$\hat{Y}_{\text{tot}}[\tau] = \hat{\beta}_{\text{tot},0} + \hat{\beta}_{\text{tot},1} X[\tau] + \hat{\beta}_{\text{tot},2} \hat{Y}_1[\tau] + \hat{\beta}_{\text{tot},3} \hat{Y}_2[\tau].$$

Remark 3. The regularization in (1.10) can be implemented using standard LASSO software by reparametrizing in terms of $\beta'_{\text{tot}} = (\beta_{\text{tot},0}, \beta_{\text{tot},1}, \beta_{\text{tot},2} - 1, \beta_{\text{tot}_3} - 1)$ and subtracting $\hat{Y}_1[t]$ and $\hat{Y}_2[t]$ from the observation of $Y_{\text{tot}}[t]$ before fitting the model. This gives estimates $\hat{\beta}'_{\text{tot}} = (\hat{\beta}'_{\text{tot},0}, \hat{\beta}'_{\text{tot},1}, \hat{\beta}'_{\text{tot},2}, \hat{\beta}'_{\text{tot},3})$ for β'_{tot}, which we turn back into estimates $\hat{\beta}_{\text{tot}} = (\hat{\beta}'_{\text{tot},0}, \hat{\beta}'_{\text{tot},1}, \hat{\beta}'_{\text{tot},2} + 1, \hat{\beta}'_{\text{tot},3} + 1)$ for β_{tot}.

Reconciliation The procedure outlined above gives us a set of forecasts $\hat{\mathbf{Y}} = (\hat{Y}_1, \hat{Y}_2, \hat{Y}_{\text{tot}})$ for any time τ, but these forecasts need not be aggregate consistent. It therefore remains to reconcile them. We will compare GTOP reconciliation to the bottom-up forecaster and to the OLS method. To apply GTOP, we have to choose the set $\mathscr{B}$, which specifies any prior knowledge we may have about the data. The easiest would be to specify no prior knowledge (by taking $\mathscr{B} = \mathbb{R}^3$), but instead we will opt to define prediction intervals for the two regional predictions as in (1.3). We will use the same prediction bounds B_1 and B_2 for the entire test set, which are estimated (somewhat simplistically) by the 95% quantile of the absolute value of the residuals in the corresponding region in the train set.

Results on the Test Set We evaluate the three reconciliation procedures bottom-up, OLS and GTOP by summing up their losses (1.4) on the test set, giving the totals L_{BU}, L_{OLS} and L_{GTOP}, which we compare to the sum of the losses $\hat{L}$ for the unreconciled forecasts by computing the *percentage of improvement* $(\hat{L} - L)/\hat{L} \times 100\%$ for $L \in \{L_{\text{BU}}, L_{\text{OLS}}, L_{\text{GTOP}}\}$. It remains to define the weighting factors a_1, a_2 and a_{tot} in the loss, and the scales σ and τ for the noise variables. We consider five different sets of weighting factors, where the first three treat the two regions symmetrically (by assigning them both weight 1), which seems the most realistic, and the other two respectively introduce a slight and a very large asymmetry between regions, which is perhaps less realistic, but was necessary to find a case where OLS would beat GTOP. Finally, we always let $\sigma + \tau = 2$, so that the scale of the noise is (somewhat) comparable between experiments. Table 1.1 shows the median over 100 repetitions of the experiment of the percentages of improvement.

First, we remark that, in all but one of the cases, GTOP reconciliation performs at least as good as or better than OLS and bottom-up, and GTOP is the only of the three methods that always improves on the unreconciled forecasts, as was already guaranteed by Theorems 1 and 2. Moreover, the only instance where OLS performs better than GTOP ($a_1 = 1, a_2 = a_{\text{tot}} = 20$), appears to be the least realistic, because the regions are treated very asymmetrically. For all cases where the weights are equal

Table 1.1 Percentage of improvement over unreconciled forecasts for simulated data

σ	τ	a_1	a_2	a_{tot}	bottom-up	OLS	GTOP
0	2	1	1	1	-13.97%	**0.40%**	**0.40%**
0	2	1	1	2	-19.47%	-2.35%	**0.47%**
0	2	1	1	10	-26.62%	-7.46%	**0.12%**
0	2	2	1	5	-22.49%	-4.55%	**0.23%**
0	2	1	20	20	-26.96%	-2.69%	**0.13%**
1	1	1	1	1	-55.51%	**5.75%**	**5.75%**
1	1	1	1	2	-75.09%	-6.02%	**4.54%**
1	1	1	1	10	-141.66%	-30.39%	**2.41%**
1	1	2	1	5	-92.47%	-14.09%	**3.13%**
1	1	1	20	20	-77.18%	-2.51%	**1.22%**
2	0	1	1	1	-94.92%	**29.85%**	**29.85%**
2	0	1	1	2	-184.23%	17.57%	**34.76%**
2	0	1	1	10	-996.22%	-79.58%	**44.75%**
2	0	2	1	5	-319.30%	1.32%	**35.48%**
2	0	1	20	20	-183.95%	**23.54%**	16.19%

($a_1 = a_2 = a_{\text{tot}} = 1$), we see that GTOP and OLS perform exactly the same, which, in light of the equivalence discussed in Section 1.2.3, suggest that the prediction intervals that make up $\mathscr{B}$ do not have a large effect in this case.

Secondly, we note that the unreconciled predictions are much better than the bottom-up forecasts. Because bottom-up and the unreconciled forecasts make the same predictions $\hat{Y}_1$ and $\hat{Y}_2$ for the two regions, this means that the difference must be in the prediction $\hat{Y}_{\text{tot}}$ for the sum of the regions, and so, indeed, the method described in (1.9) and (1.10) makes significantly better forecasts than the simple bottom-up forecast $\hat{Y}_1 + \hat{Y}_2$. We also see an overall trend that the scale of the percentages becomes larger as σ increases (or τ decreases), which may be explained by the fact that forecasting Y_{tot} becomes relatively easier, so that the difference between $\hat{Y}_{\text{tot}}$ and $\hat{Y}_1 + \hat{Y}_2$ gets bigger, and the effect of reconciliation gets larger.

1.3.2 EDF Data

To illustrate how GTOP reconciliation works on real data, we use electricity demand data provided by Électricité de France (EDF). The data are historical demand records ranging from 1 July 2004 to 31 December 2009, and are sampled each 30 minutes. The total demand is split up into $K = 17$ series, each representing a different electricity tariff. The series are divided into a calibration set (from 1 July 2004 to 31 December 2008) needed by the prediction models, and a validation set (from 1 January 2009 to the end) on which we will measure the performance of GTOP.

Every night at midnight, forecasts are required for the whole next day, i.e. for the next 48 time points. We use a non-parametric function-valued forecasting model by Antoniadis et al. [2012], which treats every day as a 48-dimensional vector. The

model uses all past data on the calibration and validation sets. For every past day d, it considers day $d+1$ as a candidate prediction and then it outputs a weighted combination of these candidates in which the weight of day d depends on its similarity to the current day. This forecasting model is used independently on each of the 17 individual series and also on the aggregate series (their total).

We now use bottom-up, OLS and GTOP to reconcile the individual forecasts. Similarly to the simulations in the previous section, the prediction intervals $B_1,\dots,B_K$ for GTOP are computed as quantiles of the absolute values of the residuals, except that now we only use the past two weeks of data from the validation set, and we use the q-th quantile, where q is a parameter. We note that, for the special case $q = 0\%$, we would expect B_k to be close to 0, which makes GTOP very similar to the bottom-up forecaster. (See Example 2.)

For each of the three methods, the percentages of improvement on the validation set are computed in the same way as in the simulations in the previous section. Table 1.2 shows their values for different choices of realistic weighting factors, using $q = 10\%$ for GTOP, which was found by optimizing for the weights $a_{\text{tot}} = 17$ and $a_k = 1$ $(k = 1,\dots,17)$, as will be discussed below.

Table 1.2 Percentage of improvement over unreconciled forecasts for EDF data, using $q = 10\%$ for GTOP

a_1	a_2	a_{tot}	bottom-up	OLS	GTOP
1	1	1	0.98%	0.19%	**1.62%**
1	1	2	1.27%	0.27%	**1.96%**
1	1	10	1.65%	0.38%	**2.41%**
1	1	17	1.70%	0.40%	**2.47%**

We see that GTOP consistently outperforms both the bottom-up and the OLS predictor, with gains that increase with a_{tot}. Unlike in the simulations, however, the bottom-up forecaster is comparable to or even better than the unreconciled forecasts in terms of its percentage of improvement. In light of Remark 2, we have therefore considered simply replacing our prediction for the total by the bottom-up predictor, which would make reconciliation unnecessary. However, when, instead of looking at the percentage of improvement, we count the times when the unreconciled forecaster gives a better prediction for the total than the bottom-up forecaster, we see that this is 56% percent, so the unreconciled forecaster does predict better than bottom-up slightly more than half of the time, and consequently there is something to gain by using it. As will be discussed next, this does make it necessary to use a small quantile q with GTOP.

Choosing the Quantile To determine which quantile q to choose for GTOP, we plot its percentage of improvement as a function of q for the case $a_{\text{tot}} = 17$ and $a_k = 1$ (see Figure 1.4). We see that all values below 60% improve on the bottom-up forecaster, and that any value below 30% improves on OLS. The quantile $q \approx 10\%$ gives the best results, and, for ease of comparison, we use this same value in all

the experiments reported in Table 1.2. In light of the interpretation of the prediction intervals, it might appear surprising that the optimal value for q would be so small. This can be explained by the fact that the unreconciled forecasts are only better than bottom-up 56% of the time, so that a small value of q is beneficial, because it keeps the GTOP forecasts close to the bottom-up ones.

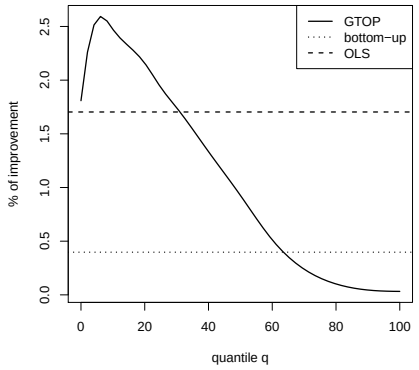

Fig. 1.4 Percentages of improvement as a function of q for GTOP, OLS and bottom-up, using $a_{\text{tot}} = 17$ and $a_k = 1$ ($k = 1, \ldots, 17$)

1.4 Discussion

We now turn to several subjects that we have not been able to treat in full detail in the previous parts of the paper. First, in Section 1.4.1, we discuss appropriate choices for the weighting factors that determine the loss. Then, in Section 1.4.2, we discuss how estimating proportions in top-down forecasting is complicated by the presence of independent variables, and, finally, in Section 1.4.3, we conclude with a summary of the paper and directions for future work.

1.4.1 How to Choose the Weighting Factors in the Loss

In the General Forecasting Competition 2012 [Hong et al., 2013], a two-level hierarchy was considered with weights chosen as $a_k = 1$ for $k = 1, \ldots, K$ and $a_{\text{tot}} = K$, so that the forecast for the total receives the same weight as all the regional forecasts taken together. At first sight this appears to make sense, because predicting the total is more important than predicting any single region. However, one should

also take into account the fact that the errors in the predictions for the total are on a much larger scale than the errors in the predictions for the regions, so that the total is already a dominant factor in the loss without assigning it a larger weight.

To make this argument more precise, let us consider a simplified setting in which we can compute expected losses. To this end, define random variables $\varepsilon_k = Y_k - \hat{Y}_k$ for the regional prediction errors at time τ and assume that, conditionally on all prior observations, 1) $\varepsilon_1, \ldots, \varepsilon_K$ are uncorrelated; and 2) the regional predictions are unbiased, so that $\mathbb{E}\{\varepsilon_k\} = 0$. Then the expected losses for the regions and the total are

$$\mathbb{E}\,\ell_k(Y_k, \hat{Y}_k) = a_k\,\mathbb{E}\left\{(Y_k - \hat{Y}_k)^2\right\} = a_k \operatorname{Var}(\varepsilon_k) \qquad (k = 1, \ldots, K)$$

$$\mathbb{E}\,\ell_{\text{tot}}(Y_{\text{tot}}, \hat{Y}_{\text{tot}}) = a_{\text{tot}}\,\mathbb{E}\left\{(\sum_k Y_k - \sum_k \hat{Y}_k)^2\right\} = a_{\text{tot}} \operatorname{Var}\Big(\sum_{k=1}^{K} \varepsilon_k\Big) = a_{\text{tot}} \sum_{k=1}^{K} \operatorname{Var}(\varepsilon_k),$$

where $\operatorname{Var}(Z)$ denotes the variance of a random variable Z.

We see that, even without assigning a larger weight to the total, $\mathbb{E}\,\ell_{\text{tot}}(Y_{\text{tot}}, \hat{Y}_{\text{tot}})$ is already of the same order as the sum of all $\mathbb{E}\,\ell_k(Y_k, \hat{Y}_k)$ together, which suggests that choosing a_{tot} to be 1 or 2 (instead of K) might already be enough to assign sufficient importance to the prediction of the total.

1.4.2 The Limits of Top-Down Forecasting

As a thought experiment, think of a noiseless situation in which

$$Y_1[t] = X[t], \qquad Y_2[t] = X[t] + 1, \qquad Y_{\text{tot}}[t] = Y_1[t] + Y_2[t] = 2X[t] + 1$$

for some independent variable $(X[t])$. Suppose we use the following top-down approach: first we estimate $Y_{\text{tot}}[\tau]$ by $\hat{Y}_{\text{tot}}[\tau]$ and then we make regional forecasts as $\hat{Y}_1[\tau] = \lambda \hat{Y}_{\text{tot}}[\tau]$ and $\hat{Y}_2[\tau] = (1-\lambda)\hat{Y}_{\text{tot}}[\tau]$ according to a constant λ that we will estimate. Because we are in a noise-free situation, let us assume that estimation is easy, and that we can predict $Y_{\text{tot}}[\tau]$ exactly: $\hat{Y}_{\text{tot}}[\tau] = Y_{\text{tot}}[\tau]$. Moreover, we will assume we can choose λ optimally as well. Then how should λ be chosen? We want to fit:

$$\lambda = \frac{Y_1[t]}{Y_{\text{tot}}[t]} = \frac{1}{2} - \frac{1}{4X[t]+2}, \qquad 1-\lambda = \frac{Y_2[t]}{Y_{\text{tot}}[t]} = \frac{1}{2} + \frac{1}{4X[t]+2}.$$

But now we see that the optimal value for λ depends on $X[t]$, which is not a constant over time! So estimating λ based on historical proportions will not work in the presence of independent variables.

1.4.3 Summary and Future Work

Unlike previous approaches, like bottom-up, top-down and generalized least-squares forecasting, we propose to split the problem of hierarchical time series forecasting into two parts: first one constructs the best possible forecasts for the time series without worrying about aggregate consistency or theoretical restrictions like unbiasedness, and then one uses the GTOP reconciliation method proposed in Section 1.2 to turn these forecasts into aggregate consistent ones. As shown by Theorems 1 and 2, GTOP reconciliation can only make any given set of forecasts better, and the less consistent the given forecasts are, the larger the improvement guaranteed by GTOP reconciliation.

Our treatment is for the squared loss only, but, as pointed out in Section 1.2, Theorems 1 and 2 readily generalize to any other loss that is based on a Bregman divergence, like for example the Kullback-Leibler divergence. It would be useful to work out this generalization in detail, including the appropriate choice of optimization algorithm to compute the resulting Bregman projection.

In the experiments in Section 1.3, we have proposed some new methods for coming up with the initial forecasts, but although they demonstrate the benefits of GTOP reconciliation, these approaches are still rather simple. In future work, it would therefore be useful to investigate more advanced ways of coming up with initial forecasts, which allow for even more information to be shared between different time series. For example, it would be natural to use a Bayesian approach to model regions that are geographically close as random instances of the same distribution on regions.

Finally, there seems room to do more with the prediction intervals for the GTOP reconciled predictions as defined in (1.3). It would be interesting to explore data-driven approaches to constructing these intervals, like for example those proposed by Antoniadis et al. [2013].

Acknowledgements The authors would like to thank Mesrob Ohannessian for useful discussions, which led to the closed-form solution for the GTOP predictions in Example 1. We also thank two anonymous referees for useful suggestions to improve the presentation. This work was supported in part by NWO Rubicon grant 680-50-1112.

References

A. Antoniadis, X. Brossat, J. Cugliari, and J.-M. Poggi. Prévision d'un processus à valeurs fonctionnelles en présence de non stationnarités. Application à la consommation d'électricité. *Journal de la Société Française de Statistique*, 153(2):52–78, 2012.

A. Antoniadis, X. Brossat, J. Cugliari, and J. M. Poggi. Une approche fonctionnelle pour la prévision non-paramétrique de la consommation d'électricité. Technical Report oai:hal.archives-ouvertes.fr:hal-00814530, Hal, Avril 2013. URL http://hal.archives-ouvertes.fr/hal-00814530.

C. E. Borges, Y. K. Penya, and I. Fernández. Evaluating combined load forecasting in large power systems and smart grids. *IEEE Transactions on Industrial Informatics*, 9(3):1570–1577, 2013.

R. P. Byron. The estimation of large social account matrices. *Journal of the Royal Statistical Society, Series A*, 141:359–367, 1978.

N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

B. Chen. A balanced system of industry accounts for the U.S. and structural distribution of statistical discrepancy. Technical report, Bureau of Economic Analysis, 2006. URL http://www.bea.gov/papers/pdf/reconciliation_wp.pdf.

G. Fliedner. An investigation of aggregate variable time series forecast strategies with specific subaggregate time series statistical correlation. *Computers & Operations Research*, 26:1133–1149, 1999.

C. W. J. Granger. Aggregation of time series variables — a survey. Discussion Paper 1, Federal Reserve Bank of Minneapolis, Institute for Empirical Macroeconomics, 1988. URL http://www.minneapolisfed.org/research/DP/DP1.pdf.

E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2–3):169–192, 2007.

T. Hong, P. Pinson, and S. Fan. Global energy forecasting competition 2012. *International Journal of Forecasting*, 2013. To appear.

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

M. S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret. Applications of second-order cone programming. *Linear Algebra and its Applications*, 284(1–3):193–228, 1998.

H. Lütkepohl. Forecasting aggregated time series variables: A survey. Working Paper EUI ECO: 2009/17, European University Institute, 2009. URL http://hdl.handle.net/1814/11256.

J. G. MacKinnon and H. White. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29:305–325, 1985.

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

R. Stone, D. G. Champernowne, and J. E. Meade. The precision of national income estimates. *The Review of Economic Studies*, 9(2):111–125, 1942.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

H. White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838, 1980.