Regularization in Hierarchical Time Series Forecasting With Application to Electricity Smart Meter Data

Souhaib Ben Taieb
Monash Business School
Monash University

Jiafan Yu
Dept. of CEE
Stanford University

Mateus Neves Barreto
FEEC
Universidade Estadual de Campinas

Ram Rajagopal
Dept. of CEE
Stanford University

Abstract

Accurate electricity demand forecast plays a key role in sustainable power systems. It enables better decision making in the planning of electricity generation and distribution for many use cases. The electricity demand data can often be represented in a hierarchical structure. For example, the electricity consumption of a whole country could be disaggregated by states, cities, and households. Hierarchical forecasts require not only good prediction accuracy at each level of the hierarchy, but also the consistency between different levels. State-of-the-art hierarchical forecasting methods usually apply adjustments on the individual level forecasts to satisfy the aggregation constraints. However, the high-dimensionality of the unpenalized regression problem and the estimation errors in the high-dimensional error covariance matrix can lead to increased variability in the revised forecasts with poor prediction performance. In order to provide more robustness to estimation errors in the adjustments, we present a new hierarchical forecasting algorithm that computes sparse adjustments while still preserving the aggregation constraints. We formulate the problem as a high-dimensional penalized regression, which can be efficiently solved using cyclical coordinate descent methods. We also conduct experiments using a large-scale hierarchical electricity demand data. The results confirm the effectiveness of our approach compared to state-of-the-art hierarchical forecasting methods, in both the sparsity of the adjustments and the prediction accuracy. The proposed approach to hierarchical forecasting could be useful for energy generation including solar and wind energy, as well as numerous other applications.

Introduction

Recently, massive amounts of detailed individual electricity consumption data has been collected by newly deployed smart meters in households (Zheng, Gao, and Lin 2013). A key technical challenge nowadays is to analyze such data to better predict the electricity generation and demand to achieve a more sustainable future (Ramchurn et al. 2012).

For example, demand forecasting is critical in understanding the effect of demand response actions (Siano 2014). There is a rich literature on forecasting the aggregated electricity demand (Hong 2010; Ba et al. 2012; Ben Taieb and Hyndman 2014). The literature on forecasting the smart meter electricity demand is much more sparse (Mirowski et al. 2014).

We focus on electricity demand forecasting where the data is represented in a hierarchical structure. The most disaggregated level usually contains a huge amount of time series, and there are additional series at higher levels of aggregation. The various time series of a hierarchy can interact in varying and complex ways. In particular, time series at different levels of the hierarchy can contain very different patterns. For example, electricity demand at the bottom level is typically very noisy sometimes exhibiting intermittency, while aggregated demand at higher levels is much smoother.

When forecasting these hierarchical structures, we must ensure that the forecasts satisfy the aggregation constraints, i.e. the forecasts must add up consistently between levels of aggregation. The simplest approach to generate hierarchical forecasts entails summing the individual bottom series forecasts. However, the high level of noise at the bottom level and the error correlations often imply poor prediction accuracy at upper levels in the hierarchy. Instead, it is also possible to forecast all series of the hierarchy independently, but it has the undesirable property of not satisfying the aggregation constraints.

Hyndman et al. (2011) proposed to compute revised forecasts from independent forecasts of different levels to match the aggregation constraints. Their approach is based on optimal forecasts combination obtained by solving a linear regression problem using information of the error forecast covariances. Williams et al. (2016) used similar regression-based forecasts combination to adaptively combine temperature forecasts in a non-hierarchal setting. Mijung and Marcel (2014) proposed a top-down forecasting method based on forecast disaggregation in a Bayesian framework. However, Hyndman et al. (2011) showed that any top-down method introduces bias in the forecasts even if the base forecasts are unbiased.

By making strong assumptions on the covariance structure of the forecast errors, Hyndman et al. (2011) computed the revised forecasts using ordinary least squares. Extensions to the method of Hyndman et al. (2011) have been proposed in Hyndman, Lee, and Wang (2016) and Wickramasuriya, Athanasopoulos, and Hyndman (2015) with relaxed assumption on the error covariance matrix and better com-

putational complexity.

In all these algorithms, the revised forecasts are obtained by adding an adjustment to the base forecasts in order to satisfy the aggregation constraints at upper levels in the hierarchy. However, the high-dimensionality of the unpenalized regression and the estimation errors in the high-dimensional error covariance matrix can lead to increased variability in the revised forecasts with poor prediction performance.

We propose a new hierarchical forecasting algorithm that seeks the sparsest adjustments in the base forecasts while still preserving aggregate consistency. The algorithm is based on a high-dimensional generalized elastic net regression problem. With such formulation, it allows some base forecasts to remain unchanged, which can lead to more robustness to estimation errors in the adjustments. Another advantage of the proposed method is the ability to dynamically switch between the simple bottom-up forecasts and the best linear unbiased revised forecasts. Finally, by including an additional constraint, the algorithm can also guarantee the non-negativity of the revised forecasts, which is particularly important when dealing with non-negative quantities such as electricity demand.

Hierarchical Time Series Forecasting

A hierarchical time series is a multivariate time series with an hierarchical structure (Fliedner 2001). We let $\boldsymbol{y}_t$ be an n-vector containing all observations at time t, and $\boldsymbol{b}_t$ be an n_b-vector with the observations at the bottom level only. We can then write

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

where $\boldsymbol{S} \in \mathbb{R}^{n_b \times n}$ is a summing matrix. We can equivalently write

$$\begin{bmatrix} \boldsymbol{a}_t \\ \boldsymbol{b}_t \end{bmatrix} = \begin{bmatrix} \boldsymbol{S}_a \\ \boldsymbol{I}_{n_b} \end{bmatrix} \boldsymbol{b}_t, \tag{2}$$

where $\boldsymbol{a}_t$ is an n_a-vector with the observations at the different levels of aggregation, and $\boldsymbol{I}_{n_b} \in \mathbb{R}^{n_b \times n_b}$ is an identity matrix.

Given T historical observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T$, the optimal h-period ahead forecasts under mean squared error (MSE) loss are given by the conditional mean, i.e.

$$\mathbb{E}[\boldsymbol{y}_{T+h}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T] = \boldsymbol{S}\,\mathbb{E}[\boldsymbol{b}_{T+h}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T], \tag{3}$$

where $h = 1, 2, \ldots, H$.

Expression (3) suggests a plug-in estimator to obtain mean forecasts for $\boldsymbol{y}_{T+h}$, i.e for all series in the hierarchy. More precisely, we can compute

$$\hat{\boldsymbol{y}}_{T+h} = \boldsymbol{S}\hat{\boldsymbol{b}}_{T+h}, \tag{4}$$

where $\hat{\boldsymbol{b}}_{T+h}$ are mean forecasts for $\boldsymbol{b}_{T+h}$. These forecasts are known as *bottom-up* (BU).

In practice, the series at the most disaggregated level are hard to predict notably due to low signal-to-noise ratio. As a result, the bottom-up forecasts will provide poor forecast accuracy at upper levels in the hierarchy.

Another method called *top-down* consists in forecasting the completely aggregated series, and then using historical proportions to disaggregate the forecasts. Compared with the bottom-up method, this method has the advantage of computing forecasts for a time series with a high signal-to-noise ratio. However, the forecast accuracy will depend on the accuracy of the disaggregation procedure, and all the information at the intermediate and bottom levels is disregarded.

Instead of aggregating or disaggregating forecasts, it is also possible to forecast all series at all levels independently, i.e. estimating each component on the the left hand side of expression (3). In other words, we compute

$$\hat{\boldsymbol{y}}_{T+h} = \begin{bmatrix} \hat{\boldsymbol{a}}_{T+h} \\ \hat{\boldsymbol{b}}_{T+h} \end{bmatrix}, \tag{5}$$

which we call the *base* forecasts (BASE).

This brings a lot of flexibility since we can use different methods at different levels, e.g. advanced forecasting methods at higher levels, to capture smooth patterns and simple methods at bottom levels to accommodate high signal-to-noise ratio.

However, due to forecast errors, the base forecasts will not satisfy the aggregation constraints. We define the consistency error as

$$\tilde{\boldsymbol{\varepsilon}}_{T+h} = \hat{\boldsymbol{a}}_{T+h} - \boldsymbol{S}_a\hat{\boldsymbol{b}}_{T+h}. \tag{6}$$

If $\tilde{\boldsymbol{\varepsilon}}_{T+h} = \boldsymbol{0}$, the hierarchical forecasts $[\hat{\boldsymbol{a}}_{T+h}\ \hat{\boldsymbol{b}}_{T+h}]'$ enjoy the *aggregate consistency* property.

Since the optimal forecasts given in (3) are aggregate consistent by definition, it is necessary to impose aggregate consistency when generating hierarchal forecasts. Also, from a decision making perspective, this property guarantees consistent decisions over the entire hierarchy.

Hyndman et al. (2011) observed that all existing hierarchical forecasting methods can be written as

$$\tilde{\boldsymbol{y}}_{T+h} = \boldsymbol{S}\boldsymbol{P}\hat{\boldsymbol{y}}_{T+h}, \tag{7}$$

for some appropriately chosen matrix $\boldsymbol{P} \in \mathbb{R}^{n_b \times n}$, and where $\hat{\boldsymbol{y}}_{T+h}$ is the base forecast, and $\tilde{\boldsymbol{y}}_{T+h}$ is the revised forecasts satisfying the aggregation constraints.

In other words, existing methods compute a linear combination of the base forecasts to obtain revised bottom forecasts, which are then summed by $\boldsymbol{S}$ to obtain the final forecasts for the whole hierarchy.

For example, we obtain bottom-up forecasts with $\boldsymbol{P} = [\boldsymbol{0}_{n_a \times n_b}|\boldsymbol{1}_{n_b \times n_b}]$, and top-down forecasts with $\boldsymbol{P} = \left[\boldsymbol{p}_{n_b \times 1}|\boldsymbol{0}_{n_b \times (n-1)}\right]$ where $\boldsymbol{p}$ is a vector of proportion that sums to one. Of course, variations are possible by defining the matrix $\boldsymbol{P}$ appropriately.

Best Linear Unbiased Revised Forecasts

One approach to find optimal $\boldsymbol{P}$ is to minimize the mean square forecast errors, i.e.

$$\underset{\boldsymbol{P}}{min.}\ \mathbb{E}\left[\left\|\boldsymbol{y}_{T+h} - \tilde{\boldsymbol{y}}_{T+h}\right\|_2^2\right], \tag{8}$$

where $\tilde{\boldsymbol{y}}_{T+h}$ is defined in (7).

Assuming unbiased base forecasts, Wickramasuriya, Athanasopoulos, and Hyndman (2015) showed that (8) admits a closed-form solution for the best (i.e. having minimum variance) linear unbiased (BLU) revised forecasts. The

solution is given by

$$\boldsymbol{P}^* = (\boldsymbol{S}'\boldsymbol{W}_h^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}_h^{-1}, \tag{9}$$

where $\boldsymbol{W}_h$ is the positive definite covariance matrix of the h-period ahead base forecast errors, i.e.

$$\boldsymbol{W}_h = \mathbb{E}[\hat{\boldsymbol{e}}_{T+h}\hat{\boldsymbol{e}}'_{T+h}] = \begin{bmatrix} \boldsymbol{W}_{h,\text{a}} & \boldsymbol{W}_{h,\text{ab}} \\ \boldsymbol{W}'_{h,\text{ab}} & \boldsymbol{W}_{\text{h,b}} \end{bmatrix}, \tag{10}$$

where $\hat{\boldsymbol{e}}_{T+h} = \boldsymbol{y}_{T+h} - \hat{\boldsymbol{y}}_{T+h}$. We will denote this method `MinT`.

Computing the matrix $\boldsymbol{P}^*$ using (9) involves the inversion of an $n \times n$ matrix where n is the total number of series in the hierarchy. By exploiting the particular structure of the summing matrix $\boldsymbol{S}$, we can show that

$$\boldsymbol{P}^* = [\boldsymbol{P}_{\tilde{\boldsymbol{\varepsilon}}} \quad \boldsymbol{I} - \boldsymbol{P}_{\tilde{\boldsymbol{\varepsilon}}}\boldsymbol{S}_a], \tag{11}$$

where

$$\begin{aligned} \boldsymbol{P}_{\tilde{\boldsymbol{\varepsilon}}} &= (\boldsymbol{W}_{h,b}\boldsymbol{S}'_a - \boldsymbol{W}'_{h,ab}) \\ &\times [\boldsymbol{W}_{h,a} - \boldsymbol{S}_a\boldsymbol{W}'_{h,ab} + (\boldsymbol{S}_a\boldsymbol{W}_{h,b} - \boldsymbol{W}_{h,ab})\boldsymbol{S}'_a]^{-1}, \end{aligned} \tag{12}$$

which only requires the inversion of an $n_a \times n_a$ matrix, where n_a is the number of series at the aggregation levels, which is orders of magnitude smaller than n in practice.

In summary, the `MinT` revised forecasts for the entire hierarchy can be computed as

$$\tilde{\boldsymbol{y}}_{T+h} \overset{\text{def}}{=} \boldsymbol{S}\tilde{\boldsymbol{b}}_{T+h}, \tag{13}$$

where $\tilde{\boldsymbol{b}}_{T+h}$ are the `MinT` bottom revised forecasts given by

$$\tilde{\boldsymbol{b}}_{T+h} \overset{\text{def}}{=} \boldsymbol{P}^*\hat{\boldsymbol{y}}_{T+h}. \tag{14}$$

Estimation of the Forecast Error Covariance

The BLU revised forecasts in (14) depend on the error covariance matrix $\boldsymbol{W}_h$ as can be seen in (9). However, in practice, this matrix is not available and needs to be estimated using historical observations of the base forecast errors. Since $\boldsymbol{W}_h$ is particularly challenging to estimate especially for $h > 1$, we will assume $\boldsymbol{W}_h = k_h\boldsymbol{W}_1$ as in Wickramasuriya, Athanasopoulos, and Hyndman (2015). Then, the matrix can be estimated via

$$\hat{\boldsymbol{W}}_1 = \kappa\frac{\sum_{t=1}^{T}\lambda^{T-t}\hat{\boldsymbol{e}}_t\hat{\boldsymbol{e}}'_t}{\sum_{t=1}^{T}\lambda^{T-t}}, \tag{15}$$

where κ is a normalizing constant and λ $(0 < \lambda \leq 1)$ is an exponential time decay parameter, which allows more emphasis to be placed on the more recent observations. If $\kappa = \frac{T}{T-1}$ and $\lambda = 1$, then the estimate reduces to the sample covariance.

However, since in practice the matrix $\boldsymbol{W}_1$ is high-dimensional, i.e. $T = \mathcal{O}(n)$, the sample covariance estimator will perform poorly notably due to the accumulation of estimation errors. In particular, the sample covariance matrix will not always be positive definite, and hence not invertible.

To overcome this limitation, one can make structural assumptions on the entries of the matrix. For example, Hyndman et al. (2011) used $\hat{\boldsymbol{W}}_{1,\text{OLS}} = \boldsymbol{I}$, the most simplifying assumption. Hyndman, Lee, and Wang (2016) used $\hat{\boldsymbol{W}}_{1,\text{WLS}} = \text{diag}(\hat{w}_1, \ldots, \hat{w}_n)$ where $\hat{w}_j$ is an estimate of the base forecast error variance for series $j = 1, 2, \ldots, n$.

In this work, we will consider a shrinkage estimator with a diagonal target given by

$$\hat{\boldsymbol{W}}_{1,\text{shrink}} = \lambda\hat{\boldsymbol{W}}_{1,\text{WLS}} + (1-\lambda)\hat{\boldsymbol{W}}_1, \tag{16}$$

where $\lambda \in [0, 1]$ is a shrinkage factor. As a result, the off-diagonal entries of the matrix are shrunk towards zero, while the diagonal entries remains unchanged. A closed-form expression for the optimal shrinkage factor has been proposed by Schäfer and Strimmer (2005), and is given by

$$\hat{\lambda} = \frac{\sum_{i\neq j}\hat{\text{Var}}(\hat{r}_{ij})}{\sum_{i\neq j}\hat{r}_{ij}^2} \tag{17}$$

where $\hat{r}_{ij}$ is the ij-th entry of the sample correlation matrix $\hat{\boldsymbol{R}}_1$. The same shrinkage estimator has been used in Wickramasuriya, Athanasopoulos, and Hyndman (2015) for hierarchical forecasting with tourism applications.

Regularization in Hierarchical Forecasting

The `MinT` bottom revised forecasts given by (14) can also be computed as the Generalised Least Squares (GLS) solution of the following regression model:

$$\hat{\boldsymbol{y}}_{T+h} = \boldsymbol{S}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{T+h}, \tag{18}$$

where $\boldsymbol{\beta} = \mathbb{E}[\boldsymbol{b}_{T+h}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T]$ and $\boldsymbol{\varepsilon}_{T+h} \sim \mathcal{N}(\boldsymbol{0}, \hat{\boldsymbol{W}}_h)$.

The previous regression model is high-dimensional since it involves $n = n_a + n_b$ observations and n_b predictors with $n = \mathcal{O}(n_b)$ where n_a and n_b are the number of series in the aggregation levels and the bottom level, respectively.

If we apply the change of variable $\boldsymbol{\beta} = \hat{\boldsymbol{b}}_{T+h} + \boldsymbol{\theta}$, then the optimization problem of the GLS regression can be formulated as

$$\underset{\boldsymbol{\theta}\in R^{n_b}}{min.}\ (\hat{\boldsymbol{z}}_{T+h} - \boldsymbol{S}\boldsymbol{\theta})'\hat{\boldsymbol{W}}_h^{-1}(\hat{\boldsymbol{z}}_{T+h} - \boldsymbol{S}\boldsymbol{\theta}), \tag{19}$$

where

$$\hat{\boldsymbol{z}}_{T+h} = \hat{\boldsymbol{y}}_{T+h} - \boldsymbol{S}\hat{\boldsymbol{b}}_{T+h} = [\tilde{\boldsymbol{\varepsilon}}_{T+h} \quad \boldsymbol{0}]', \tag{20}$$

and $\tilde{\boldsymbol{\varepsilon}}_{T+h}$ is the consistency error defined in (6). The bottom revised forecasts are then given by

$$\tilde{\boldsymbol{b}}_{T+h} = \hat{\boldsymbol{b}}_{T+h} + \hat{\boldsymbol{\theta}}, \tag{21}$$

where $\hat{\boldsymbol{\theta}}$ is the estimated adjustment applied to the base forecasts.

Furthermore, if we plug (11) in (14), we can see that the `MinT` adjustment has a closed-form expression given by

$$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{P}}_{\tilde{\boldsymbol{\varepsilon}}}\,\tilde{\boldsymbol{\varepsilon}}_{T+h}. \tag{22}$$

The previous expression shows the importance of properly estimating $\hat{\boldsymbol{W}}_h$ since the adjustment $\hat{\boldsymbol{\theta}}$ depends on $\hat{\boldsymbol{P}}_{\tilde{\boldsymbol{\varepsilon}}}$, which is computed using $\hat{\boldsymbol{W}}_h$ as can be seen in (12). In other words, the estimation errors in the high-dimensional matrix $\hat{\boldsymbol{W}}_h$ could negatively affect the revised forecasts. In particular, when $\tilde{\boldsymbol{\varepsilon}}_{T+h} \neq \boldsymbol{0}$, i.e. when the forecasts are not aggregate consistent, the `MinT` adjustment $\hat{\boldsymbol{\theta}}$ will not be sparse

even in the extreme scenario where $\hat{\boldsymbol{W}}_h$ is an identity matrix. This implies that the MinT adjustments do not allow the base forecasts to remain unchanged, i.e. $\hat{\theta}_j \neq 0$ for all j.

Finally, using (21) and (22), we can see that n_a consistency errors in $\tilde{\boldsymbol{\varepsilon}}_{T+h}$ will affect all the n_b bottom base forecasts $\tilde{\boldsymbol{b}}_{T+h}$ where n_a is significantly smaller than n_b in almost all applications of hierarchical forecasting. However, for example in hierarchal electricity demand forecasting, we want to avoid few consistency errors to affect the forecasts of every single household in the hierarchy.

Revised Forecasts With Sparse Adjustments

In order to mitigate the effect of estimation errors in the adjustments, we propose to compute the sparsest adjustments with the best forecast accuracy. It is important to note that we do not seek sparse revised forecasts but rather sparse adjustments. This is important when dealing with strictly positive demand as is often the case with electricity demand.

To preserve aggregate consistency, our revised forecasts will have the same form as (13), i.e. the revised forecasts for the entire hierarchy are computed by summing revised bottom forecasts. Sparsity can be achieved by constraining the number of non-zero coefficients $s = \|\boldsymbol{\theta}\|_0$ where $\|\cdot\|_0$ is the L_0 "norm". The MinT forecasts typically have $s = n_b$, while the BU forecasts have $s = 0$, i.e. no adjustments. Our goal is to find a good tradeoff between these two extreme cases.

Because the L_0 "norm" function is non-convex, convex relaxation such as the LASSO (Tibshirani 2011) are typically used to obtain sparse estimate in linear regression (Hastie, Tibshirani, and Wainwright 2015). We propose the following penalized (generalized) least squares problem:

$$\underset{\boldsymbol{\theta}}{min.}\ (\hat{\boldsymbol{z}}_{T+h} - \boldsymbol{S}\boldsymbol{\theta})'\hat{\boldsymbol{W}}_h^{-1}(\hat{\boldsymbol{z}}_{T+h} - \boldsymbol{S}\boldsymbol{\theta}) + \lambda \sum_{j=1}^{n_b} \gamma_j R_\alpha(\theta_j) \tag{23}$$

$$\text{s.t. } \boldsymbol{\theta} \succcurlyeq -\hat{\boldsymbol{b}}_{T+h}, \tag{24}$$

where $\lambda \geq 0$ is a penalty parameter, $\gamma_j \geq 0$ is a penalty scaling parameter,

$$R_\alpha(\theta_j) = \left[(1-\alpha)\frac{1}{2}\theta_j^2 + \alpha|\theta|_j\right] \tag{25}$$

is the elastic net penalty (Zou and Hastie 2005) with weighting parameter $\alpha \in [0, 1]$. The bottom revised forecasts are then given by (21). We will denote this method MinT-REG.

The constraint in (24) guarantees the non-negativity of the bottom revised forecasts since $\tilde{\boldsymbol{b}}_{T+h} \succcurlyeq \boldsymbol{0}$ if $\hat{\boldsymbol{b}}_{T+h} \succcurlyeq \boldsymbol{0}$. This constraint is important when dealing with non-negative quantities as in household electricity demand without energy generation.

The penalty R_α in (25) is a convex combination of the ridge penalty ($\alpha = 0$) and the lasso penalty ($\alpha = 1$). It allows a compromise between these two penalties, and is particularly useful in high-dimensional setting with correlated predictors where the lasso penalty can be unstable (Tibshirani 2013).

For all $\alpha < 1$, the penalty function is strictly convex, and when $\alpha \in (0, 1]$, it is singular at 0. In this work, we will consider $\alpha > 0$ to always include a lasso penalty, and allow sparse estimates.

When $\lambda = 0$, the objective function in (23) is equivalent to the objective in (19) with a lower limit on the adjustments $\boldsymbol{\theta}$. In other words, our algorithm also allows to implement the non-negative MinT forecasting algorithm. When $\alpha = 1$, we obtain the same forecasts as MinT ($\lambda = 0$) and BU ($\lambda = \infty$). By selecting the right amount of penalization, we can move from the completely dense adjustments (MinT) to the completely sparse adjustments (BU). Finally, since $\hat{\boldsymbol{\theta}} = \tilde{\boldsymbol{b}}_{T+h} - \hat{\boldsymbol{b}}_{T+h}$, a shrinkage of the adjustment $\hat{\boldsymbol{\theta}}$ towards zero is equivalent to a shrinkage of the bottom revised forecasts $\tilde{\boldsymbol{b}}_{T+h}$ towards the bottom base forecasts $\hat{\boldsymbol{b}}_{T+h}$.

The penalty we consider in (23) involves a non-uniform penalization (i.e. different penalties for each of the coefficients) where the penalty for the coefficient θ_j is $\lambda_j = \lambda\gamma_j$. This differential shrinkage is particularly important in our case since we do not want to shrink the large forecasts of the aggregation levels in the same way as the small forecasts of the bottom levels. Similarly to the adaptive lasso of Zou (2006), we use

$$\gamma_j = \frac{1}{|\hat{\theta}_j^{\text{MinT}}|^\delta} \tag{26}$$

where $\hat{\theta}_j^{\text{MinT}}$ are the MinT adjustments, and $\delta > 0$.

If $\hat{\boldsymbol{W}}_h = \boldsymbol{C}\boldsymbol{C}'$ then the generalized elastic net regression problem in (23) can be reduced to a standard elastic net problem with response variable $\boldsymbol{C}^{-1}\hat{\boldsymbol{z}}_{T+h}$ and design matrix $\boldsymbol{C}^{-1}\boldsymbol{S}$.

The elastic net problem can be efficiently solved by cyclical coordinate descent (Friedman et al. 2007). Our implementation of the standard elastic net is based on the glmnet package (Friedman, Hastie, and Tibshirani 2010) available for the R programming language (R Core Team 2015).

Hierarchical Electricity Demand Forecasting

Electricity Smart Meter Data

We use the data collected during a smart metering trial conducted across Great Britain by four energy supply companies (AECOM 2011). The data set contains half-hourly measurements of electricity consumption gathered from over $14,000$ households between January 2008 and September 2010, along with some geographic and demographic data.

We focus on the 5701 meters which do not have missing values, with data available between April 20,2009 and July 31, 2010; hence, each time series has $T = 22,464$ observations.

The hierarchy is based on the geographic data with six different levels with the following number of series per level: 1 (level 1), 5 (level 2), 13 (Level 3), 34 (level 4), 94 (level 5) and 5701 (level 6). In other words, the hierarchy is composed of $n = 5848$ series, $n_b = 5701$ bottom series and $n_a = n - n_b = 147$ series at aggregation levels. Figure 1 gives examples of weekly electricity demand at different levels of aggregation.

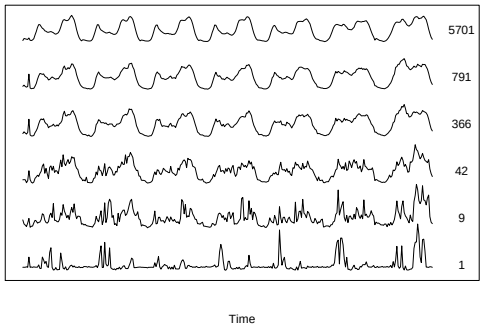

Figure 1: One week of electricity demand for different levels of aggregation, with the number of aggregated meters.

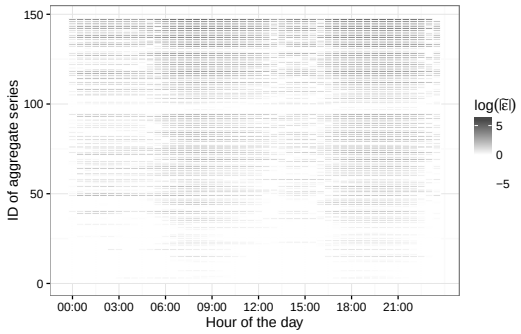

Figure 2: Absolute value of consistency errors (log. scale).

Experimental Setup

We focus on one-day ahead hierarchal electricity demand forecasting using the dataset described in the previous section. More precisely, with a forecast origin at `23:30`, we generate $H = 48$ half-hour forecasts for the next day.

We split the data into training, validation and test sets; the first 12 months for training, the next month for validation and the remaining months for testing. In both validation and testing, we generate forecasts for lead-times ranging from one-step to 48-steps ahead with `23:30` of each day as forecast origin. We end up with 1440 and 4416 observations in the training and test set, respectively.

For each time instant, we measure forecast accuracy using mean square errors computed over all series in the hierarchy. We estimate the matrix $\boldsymbol{W}_h$ using the shrinkage estimator defined in (16) with a block-diagonal target and the estimate is recomputed every 10 days in both validation and testing. Finally, a different value of α and λ in (23) is used for each half hour by minimizing mean squared forecast errors for the associated half hours in the validation set.

Forecasting Methods

We compare our forecasting algorithm `MinT-REG` given in (23) with `BU`, `BASE` and `MinT` given in (4), (5) and (13), respectively.

For the base forecasts, we use an exponential smoothing based method, called TBATS (Livera, Hyndman, and Snyder 2011), that can handle multiple seasonal cycles. This will allow use to capture the within-day and within-week seasonalities in the half-hourly demand. We model two seasonal cycles of period $m_1 = 48$ (daily) and $m_2 = 336$ (weekly). In order to stabilize the variance and guarantee the non-negativity of the base forecasts, we applied a log transformation. The parameters of the TBATS model are estimated by maximum likelihood and model selection is performed using AIC. Multi-step ahead forecasts are computed using a recursive forecasting strategy. For more details, we refer the reader to Livera, Hyndman, and Snyder (2011). Our implementation of the TBATS model is based on the forecast package (Hyndman and Khandakar 2008) available for the R programming language (R Core Team 2015).

Experimental Results

Figure 2 shows the consistency errors $\tilde{\boldsymbol{\varepsilon}}_t$ defined in (6) for the first day in the test set [1]. Each row represents one of the 147 series at upper levels in the hierarchy where the top row is associated to the most aggregated series and the other rows are ordered by levels in the hierarchy.

We can see that the consistency errors are increasing with the level of aggregation. This suggest that the `BU` forecasts is not able to capture certain patterns or seasonalities of the series at higher levels of aggregation by only using the raw bottom base forecasts. Furthermore, we can also notice a pattern by hour of the day, with lower errors during night hours compared to day hours, and higher errors for the two peak periods `06:00-12:00` and `17:00-22:00`. This can be explained by the fact that the electricity demand is typically flat during night hours, and hence easier to forecast at all levels, while the demand at peak hours is harder to forecast. The heterogeneity in the consistency errors suggest that more adjustments of the bottom base forecasts will be required at certain hours and levels of aggregation in order to properly match the base forecasts at upper levels.

Figure 3 and 4 give the MSE of the different hierarchal forecasting methods averaged over all series at the aggregation levels and all bottom series, respectively.

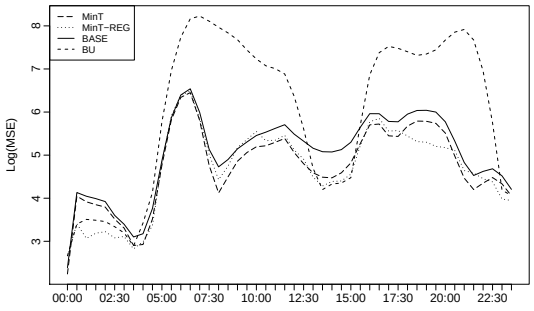

Figure 3: MSE averaged over series in the aggregation levels.

[1] Similar patterns are observed for other days in the test set.

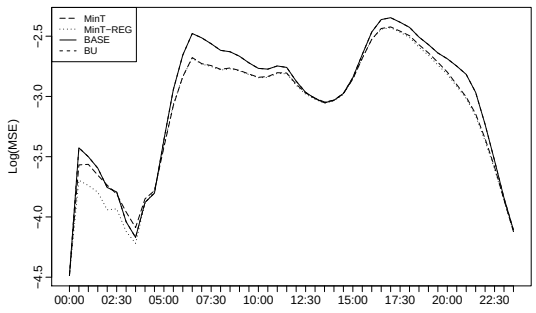

Figure 4: MSE averaged over series in the bottom level.

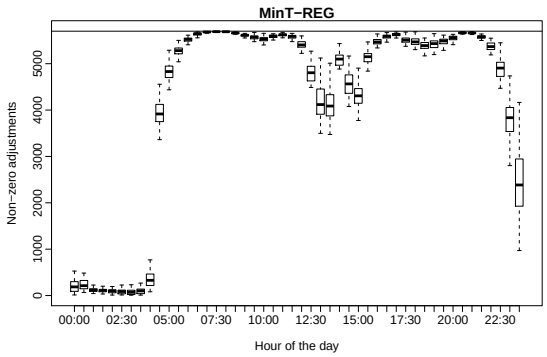

Figure 5: Number of non-zero adjustments of the MinT-REG forecasts for each hour in the test set. The horizontal line gives the maximum number of adjustments.

The solid line gives the MSE of the BASE forecasts, which are not aggregate consistent as previously show in Figure 2, and the other lines give the MSE of the aggregate consistent hierarchical forecasting methods.

In Figure 3, we can see that the BU forecasts have larger errors compared to the other methods during the two peak hours, but lower errors during night hours and around midday. This shows the limitation of computing the forecasts for the aggregation levels by summing the noisy forecasts of the bottom level.

The MinT−REG and MinT forecasts have a significantly lower MSE than the BU forecasts, and achieve similar or better performance than the BASE forecasts. The fact that both MinT−REG and MinT do not increase forecast errors compared to the BASE forecasts suggest that the adjustments applied to the bottom forecasts before summation have been properly estimated. This also shows that the bottom forecasts can benefit from the higher accuracy of the forecasts at higher levels of aggregation.

Overall, MinT−REG and MinT have similar performance except for night hours where MinT−REG has lower errors. However, MinT−REG enjoys sparser adjustments as can be seen in Figure 5. In particular, we can see the high sparsity during night hours, and the low sparsity during peak hours. In fact, MinT−REG is close to BU and MinT during night hours and peak hours, respectively. In other words, MinT−REG has the ability to dynamically switch between MinT and BU by tracking the best between these two forecasts.

The amount of sparsity in Figure 5 seems to be directly related to the amount of consistency errors in Figure 2, suggesting that less consistency errors will allow more sparsity in the adjustments. In these experiments the amount of sparsity has been automatically selected using a validation set. However, it is also possible to manually tradeoff sparsity with forecast accuracy by using the right amount of shrinkage, for example by exploiting prior knowledge about the forecasting problem.

By definition, the BASE and BU forecasts are equivalent at the bottom level, as can be seen in Figure 4. However, both MinT and MinT−REG have either similar or better forecast accuracy than BASE, which shows that the adjustments have also contributed to decrease forecast errors at the bottom level by benefiting from better forecast accuracy at upper levels in the hierarchy.

If the forecast accuracy is the only criterion, it is often hard to improve over the base forecasts, especially with a large number of series, and a lof of observations for each series, as with smart electricity demand data. However, the base forecasts do not satisfy the aggregation constraints. So, even if it is not possible to improve over the base forecasts, MinT−REG will compute sparse adjustments with the least increase in forecast errors compared to the base forecasts. By doing so, MinT−REG will provide more robustness to estimation errors in the adjustments compared to MinT.

Conclusion

By applying adjustments to the individual forecasts of an hierarchical time series, it is possible to obtain revised forecasts that satisfy the hierarchical aggregation constraints. However, in existing approaches, the computation of these adjustments involve a high-dimensional unpenalized regression and the estimation of a high-dimensional covariance matrix. As a result, the existing forecasting methods can suffer from extensively adjusted base forecasts with poor prediction accuracy.

We overcome this challenge by proposing a new forecasting algorithm that adds a sparsity constraint to the adjustments, while still preserving the aggregation constraints. The algorithm is based on a high-dimensional penalized regression with adaptive shrinkage that can be solved efficiently using cyclical coordinate descent methods.

An attractive property of the proposed method is the ability to switch in a data-driven manner between bottom-up forecasts and BLU revised forecasts. Furthermore, the sparsity of the adjustments allows substantial part of the base forecasts to remain unchanged which leads to more stability.

The experiments performed using hierarchical electricity demand data show the effectiveness of our approach compared with the state-of-the art methods. In particular, the revised forecasts enjoy a high sparsity in the adjustments during night hours, and reduce to the BLU revised forecasts during peak hours.

References

AECOM. 2011. Energy demand research project: Final analysis. Technical report, AECOM House, Hertfordshire, UK.

Ba, A.; Sinn, M.; Goude, Y.; and Pompey, P. 2012. Adaptive learning of smoothing functions: Application to electricity load forecasting. In *Advances in Neural Information Processing Systems*, 2519–2527.

Ben Taieb, S., and Hyndman, R. J. 2014. A gradient boosting approach to the kaggle load forecasting competition. *International Journal of Forecasting* 30(2):382–394.

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

Friedman, J.; Hastie, T.; Höfling, H.; and Tibshirani, R. 2007. Pathwise coordinate optimization. *The Annals of Applied Statistics* 1(2):302–332.

Friedman, J.; Hastie, T.; and Tibshirani, R. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33(1):1–22.

Hastie, T.; Tibshirani, R.; and Wainwright, M. 2015. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press.

Hong, T. 2010. *Short Term Electric Load Forecasting*. Ph.D. Dissertation.

Hyndman, R. J., and Khandakar, Y. 2008. Automatic time series forecasting: the forecast package for R.

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

Hyndman, R. J.; Lee, A. J.; and Wang, E. 2016. Fast computation of reconciled forecasts for hierarchical and grouped time series. *Computational Statistics & Data Analysis* 97:16–32.

Livera, A. M. D.; Hyndman, R. J.; and Snyder, R. D. 2011. Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association* 106(496):1513–1527.

Mijung, P., and Marcel, N. 2014. Variational bayesian inference for forecasting hierarchical time series. In *ICML 2014 Workshop on Divergence Methods for Probabilistic Inference*.

Mirowski, P.; Chen, S.; Ho, T. K.; and Yu, C.-N. 2014. Demand forecasting in smart grids. *Bell Labs Technical Journal* 18(4):135–158.

R Core Team. 2015. R: A language and environment for statistical computing.

Ramchurn, S. D.; Vytelingum, P.; Rogers, A.; and Jennings, N. R. 2012. Putting the 'smarts' into the smart grid: a grand challenge for artificial intelligence. *Communications of the ACM* 55(4):86–97.

Schäfer, J., and Strimmer, K. 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology* 4:Article32.

Siano, P. 2014. Demand response and smart grids—a survey. *Renewable and Sustainable Energy Reviews* 30:461–478.

Tibshirani, R. 2011. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society. Series B, Statistical methodology* 73(3):273–282.

Tibshirani, R. J. 2013. The lasso problem and uniqueness. *Electronic Journal of Statistics* 7:1456–1490.

Wickramasuriya, S. L.; Athanasopoulos, G.; and Hyndman, R. J. 2015. Forecasting hierarchical and grouped time series through trace minimization. Technical Report 15/15.

Williams, J. K.; Neilley, P. P.; Koval, J. P.; and McDonald, J. 2016. Adaptable regression method for ensemble consensus forecasting. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Zheng, J.; Gao, D. W.; and Lin, L. 2013. Smart meters in smart grid: An overview. In *Green Technologies Conference, 2013 IEEE*, 57–64.

Zou, H., and Hastie, T. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B, Statistical methodology* 67(2):301–320.

Zou, H. 2006. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* 101(476):1418–1429.