Forecasting in Hierarchical Models

Lucy Morgan

Supervisor: Nikolaos Kourentzes

20th February 2015

Introduction

Forecasting is the process of making statements about events whose actual outcomes (typically) have not yet been observed. It has major applications in areas such as: economics; the environment and politics, where being able to estimate future events is useful for decision making.

Often forecasting problems exhibit a natural hierarchical structure. A cross-sectional hierarchical structure is an arrangement of items in which the items are ordered above, below or on the same level as others. For example in a vehicle manufacturing company, it is useful to be able to forecast demand in order to make the right amount of vehicles. This demand may be split into different areas of production. Such a company may sell a few different types of vehicle: sports cars; off-road vehicles and vans for example and each of these can be disaggregated into finer categories involving engine size, colour or price. All of which have an effect on the demand for each vehicle. This disaggregation imposes a hierarchical structure. If the total number of vehicles sold by the company was looked at as a time series, then this could be decomposed into time series for each area of production and further within that into each vehicle colour etc. These are referred to as hierarchical time series.

This paper will look at forecasting in hierarchical models with a focus on cross-sectional hierarchies. The current methods of forecasting in hierarchical models are: top-down; bottom-up; middle-out and optimal combination. These methods will be discussed and compared with respect to their forecasting success.

Cross-sectional hierarchical models

Consider the multi-level hierarchical model shown in Figure 1 below. Here there are $K = 2$ levels. In a general hierarchical model, with K levels, level 0 is defined as the completely aggregated series. Each level from 1 to K denotes a further disaggregation down to level K containing the most disaggregated, bottom level, time series. Observations are recorded at times $t = 1, \ldots, n$ and the aim is to make forecasts for each series at each level at times $t = n + 1, \ldots, n + h$.

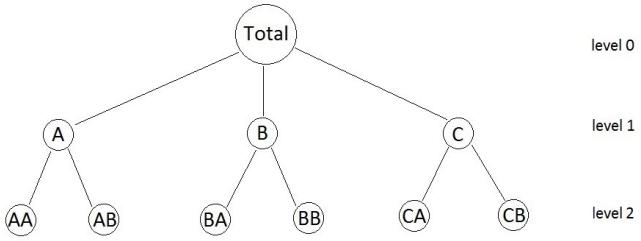

Figure 1: A two level hierarchical model.

Let Y_t be the the aggregate of all series at time t. In a hierarchical model the observations at higher levels can be obtained by summing the series below, Hyndman et al. (2011). Thus $Y_t = \sum_i Y_{i,t}$, where i indexes a generic series at level 1 of the hierarchy, and $Y_{i,t} = \sum_j Y_{ij,t}$ etc.

When considering hierarchical models it is more convenient to work with matrix and vector notation. If we let $\mathbf{Y}_{i,t}$ denote the vector of all observations at level i and time t then,

$$\mathbf{Y}_t = [Y_t, \mathbf{Y}_{1,t}, \ldots, \mathbf{Y}_{K,t}]'.$$

Using the *'summing'* matrix, $\mathbf{S}$, which stores the structure of the hierarchy, $\mathbf{Y}_t$ can be found from the bottom level series.

$$\mathbf{Y}_t = \mathbf{S}\mathbf{Y}_{K,t}$$

The *'summing'* matrix $\mathbf{S}$ is a matrix of order $m \times m_K$. Where m is the total number of series in the hierarchy and the number of series in each level is m_i (for $i = 0, 1, \ldots, K$). For the hierarchical model in Figure 1 we have,

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

The bottom-up method

The bottom-up approach, as discussed in Athanasopoulos et al. (2009), focusses on producing forecasts at the lowest level K and aggregating them to the upper levels of the hierarchy using the summing matrix $\mathbf{S}$. It works by summing together the appropriate lower level forecasts. There is no loss of information within this method so we can capture the dynamics of the individual series. But it can mean constructing a forecasting model is hard if the bottom level data is noisy, or unstructured. It also has the disadvantage of having many time series to forecast if there are many series in the lower level.

The top-down method

Athanasopoulos et al. (2009) also discuss the alternative top-down approach. Here a forecast is produced at the top level, level 0, and is then disaggregated to the lower levels of the hierarchy using proportions. The proportions themselves are usually calculated using historical data from times $t = 1, \ldots, n$. There are three typical ways of calculating these proportions described in Hyndman and Athanasopoulos (2014).

1. Average historical proportions

$$p_j = \frac{1}{n} \sum_{t=1}^{n} \frac{y_{j,t}}{y_t}$$

   Proportions are found for each series at the bottom level of the hierarchy, $j = 1, \ldots, m_K$. This method reflects the average of the historical proportions of the bottom level series relative to the total aggregate y_t, over time $t = 1, \ldots, n$.

2. Proportions of historical averages

$$p_j = \sum_{t=1}^{n} \frac{y_{j,t}}{n} / \sum_{t=1}^{n} \frac{y_t}{n}$$

   This method uses the average historical value of the bottom level series $y_{j,t}$ relative to the average value of the top level total aggregate y_t, over time $t = 1, \ldots, n$.

3. Forecasted proportions

   This method is seen as an improvement on the static proportions calculated in the previous two methods. It sees an independent base forecast generated for all series in the hierarchy. Then for each level in turn, from the top to the bottom, the proportion of each base forecast to the aggregate of all

the base forecasts at that level are calculated. These proportions are called the forecast proportions and for a hierarchy with K levels we have,

$$p_j = \prod_{l=0}^{K-1} \frac{\hat{y}_{j,t}^{(l)}}{\hat{S}_{j,t}^{(l+1)}}$$

for the bottom level series j, where $j = 1, \ldots, m_K$. Here $\hat{y}_{j,t}^{(l)}$ is the base forecast of the series that corresponds to the node which is l levels above j and $\hat{S}_{j,t}^{(l+1)}$ is the sum of the base forecasts below the series that is l levels above node j and directly in contact with that series. So for Figure 1 and the bottom level series AA,

$$\begin{aligned} p_{AA} &= \left(\frac{\hat{y}_{AA,t}}{\hat{S}_{AA,t}^{(1)}}\right)\left(\frac{\hat{y}_{AA,t}^{(1)}}{\hat{S}_{AA,t}^{(2)}}\right) \\ &= \left(\frac{\hat{y}_{AA,t}}{\hat{S}_{A,t}}\right)\left(\frac{\hat{y}_{A,t}}{\hat{S}_{Total,t}}\right) \end{aligned}$$

where $\hat{S}_{AA,t}^{(2)} = \hat{S}_{Total,t} = \hat{y}_{A,t} + \hat{y}_{B,t} + \hat{y}_{C,t}$ and $\hat{S}_{AA,t}^{(1)} = \hat{S}_{A,t} = \hat{y}_{AA,t} + \hat{y}_{AB,t}$.

Top-down forecasting has the advantage of only having to generate one forecast at the top level. It works well with low count series, but suffers from loss of information in the lower levels of the series. This means special events in individual series or trends in the data may be missed.

There is also the problem that top-down methods never give unbiased revised forecasts even when the base forecasts are unbiased. Hyndman et al. (2011) show that by letting $\hat{\mathbf{Y}}_n(h)$ denote the h-step ahead base forecast of the total aggregation then all hierarchical models can be written as,

$$\tilde{\mathbf{Y}}_n(h) = \mathbf{SP}\hat{\mathbf{Y}}_n(h).$$

Thus for some appropriately chosen matrix $\mathbf{P}$ of size $m_K \times m$ unbiasedness holds provided $\mathbf{SPS} = \mathbf{S}$. The role of $\mathbf{P}$ is to extract and combine relevant elements of the base forecasts, $\hat{\mathbf{Y}}_n(h)$, which are then summed by $\mathbf{S}$ to give the revised forecasts, $\tilde{\mathbf{Y}}_n(h)$. In top-down $\mathbf{SPS} \neq \mathbf{S}$ whatever proportion method is used, thus top-down can never give unbiased forecasts. In comparison, in the bottom-up approach $\mathbf{P} = [\mathbf{0}_{\mathbf{m_K}\times(\mathbf{m}-\mathbf{m_K})}|\mathbf{I}_{\mathbf{m_K}}]$ where $\mathbf{0}$ and $\mathbf{I}$ are respectively the null and identity matrices. Therefore $\mathbf{SPS} = \mathbf{S}$ holds for the bottom-up approach and thus, given unbiased base forecasts, it will always give unbiased revised forecasts.

The middle-out method

The middle-out method was the first extension of the top-down and bottom-up approaches. It combines ideas from both methods by starting from an intermediate (or 'middle') level, T. At this level and those below, base forecasts are calculated. From these, revised forecasts are then found for all levels. Bottom-up is used to aggregate the forecasts for those levels above T and the top-down approach is used the disaggregate the middle level forecasts downwards to the levels below T.

Optimal combination

A new approach to forecasting in hierarchical models was proposed by Hyndman et al. (2011). It incorporates the use of a regression model to optimally combine and reconcile forecasts. The method allows for correlations and interactions between series at each level to be taken into account. It is also flexible, ad hoc adjustments for each series can be made to the model allowing more information to be included for

example, special events and seasonality. The technique also has the major advantage of being able to give estimates of forecast uncertainty which allows the calculation of prediction intervals (something which was lacking from previous methods we have encountered).

Method

The method proposed by Hyndman et al. (2011) aims to estimate the unknown expectations of the future values of the bottom level of the time series, K. Let $\boldsymbol{\beta}_n(h)$ be the vector of these unknown means at the h-step ahead time point,

$$\boldsymbol{\beta}_n(h) = \mathbb{E}[\mathbf{Y}_{K,n+h}|\mathbf{Y}_1, \ldots, \mathbf{Y}_n].$$

Note: recall here $\mathbf{Y}_t$ is the vector of all observations at time t and $\mathbf{Y}_{K,n+h}$ is the vector of observations in the bottom level K at the h-step ahead time point.

Then the base forecasts $\hat{\mathbf{Y}}_n(h)$ can be written in the form of a regression equation as such,

$$\hat{\mathbf{Y}}_n(h) = \mathbf{S}\boldsymbol{\beta}_n(h)+\boldsymbol{\epsilon}_h.$$

Where the error in the regression, $\boldsymbol{\epsilon}_h$, has 0 mean and covariance matrix $\sum_h$. Using this regression equation we can obtain forecasts for all levels of the hierarchy.

A big issue in this method is having little knowledge of $\sum_h$, the covariance matrix. If $\sum_h$ was known, generalised least squares estimation could be used to obtain an unbiased estimate of $\boldsymbol{\beta}_n(h)$ which minimizes the sum of the squared residuals. In large hierarchies it may not be possible to calculate $\sum_h$ at all. To overcome this, Hyndman et al. (2011) make the assumption that the error can be estimated by the forecast error in the bottom level, $\boldsymbol{\epsilon}_h \approx \mathbf{S}\boldsymbol{\epsilon}_{K,h}$. This assumes that the errors satisfy the same aggregation constraint as the data, a reasonable assumption in most hierarchical models. Thus $\sum_h \approx \mathbf{S}\text{Var}(\boldsymbol{\epsilon}_{K,h})\mathbf{S}'$. This leads to the main result of optimal combination forecasting.

Theorem 1. Let $\mathbf{Y} = \mathbf{S}\boldsymbol{\beta}_h+ \boldsymbol{\epsilon}$ with $\text{Var}(\boldsymbol{\epsilon})=\sum_h = \mathbf{S}\text{Var}(\boldsymbol{\epsilon}_{K,h})\mathbf{S}'$ and $\mathbf{S}$ a *'summing'* matrix. Then the generalised least squares estimate of $\boldsymbol{\beta}$ is independent of $\text{Var}(\boldsymbol{\epsilon}_{K,h})$:

$$\hat{\boldsymbol{\beta}}_h = (\mathbf{S}' \textstyle\sum_h^{\dagger} \mathbf{S})^{-1}\mathbf{S}' \sum_h^{\dagger} \mathbf{Y} = (\mathbf{S}'\mathbf{S})^{-1}\mathbf{S}'\mathbf{Y}$$

with variance matrix $\text{Var}(\hat{\boldsymbol{\beta}}) = \text{Var}(\boldsymbol{\epsilon}_{K,h})$ and the Moore-Penrose pseudoinverse of $\sum_h$, denoted $\sum_h^{\dagger}$. Moreover this is the minimum variance linear unbiased estimate, (Hyndman et al., 2011).

This result eases computation of the forecasts, especially in hierarchical models with large covariance matrices. It means ordinary least squares rather than generalised least squares regression can be used for finding the revised forecasts such that,

$$\tilde{\mathbf{Y}}_n(h) = \mathbf{S}(\mathbf{S}'\mathbf{S})^{-1}\mathbf{S}'\hat{\mathbf{Y}}_n(h). \tag{1}$$

This shows the optimal combination of base forecasts is independent of the data as in (1), $\tilde{\mathbf{Y}}_n(h)$ only depend on $\mathbf{S}$ so only the structure of the hierarchy has an effect on the revised forecasts.

Also since in (1), $\mathbf{P} = (\mathbf{S}'\mathbf{S})^{-1}\mathbf{S}'$ it can easily be shown that the revised forecasts $\tilde{\mathbf{Y}}_n(h)$ are unbiased since $\mathbf{SPS} = \mathbf{S}$.

Problems

Although this method looks to be an improvement on the previous simpler methods, in practice the summing matrix $\boldsymbol{S}$ can be very large which means finding the inverse of $(\mathbf{S}'\mathbf{S})$ can be computationally prohibitive. There are three well known methods for dealing with this issue.

- This problem can be solved for relatively large hierarchies using sparse arrays saving both memory and time. It is known $\mathbf{S}$ can be represented by a sparse array as it has a low number of non-zero elements compared to its size. However for some models $\mathbf{S}$ will be so large that even this will become ineffective.

- Another method developed by Paige and Saunders (1982) uses an iterative algorithm for solving sparse linear least squares problems. This could be used for very large hierarchies with many series. Although the results are only an approximation to the actual solution the difference is negligible.

- The third method takes a different approach looking at the regression model itself. Here we have a parameter μ for each series in the hierarchy. Each μ measures the contribution of the associated level to the bottom level series below it. Looking at Figure 1 we have the parameter vector

$$\phi_n(h) = [\mu_{T,h}, \mu_{A,h}, \mu_{B,h}, \mu_{C,h}, \mu_{AA,h}, \ldots, \mu_{CB,h}]',$$

  where $\beta_{AB,h} = \mu_{T,h} + \mu_{A,h} + \mu_{AB,h}$ and the other bottom level series β_i $(i = 1, \ldots, m_K)$ values are found similarly. Therefore the regression model can be re-written

$$\hat{\mathbf{Y}}_n(h) = \mathbf{S}\mathbf{S}'\phi_n(h) + \boldsymbol{\epsilon}$$

  where $\boldsymbol{\beta}_n(h) = \mathbf{S}'\phi_n(h)$. To avoid the problem of over parametrisation zero sum constraints are added to the model such that for every split, in a series, the sum of the estimated parameters below this is 0, $\sum_i \hat{\mu}_{i,h} = 0$ and $\sum_j \hat{\mu}_{ij,h} = 0$.

  Now this re-parametrisation can be seen as an ANOVA model where the parameters are estimated by least squares. So the estimators, $\hat{\mathbf{Y}}_n(h)$, can be expressed without the need for matrix inversion, thus cutting out the computationally expensive part of the previous formulation. A drawback of this method is for unbalanced hierarchies, where on at least one level series have an unequal number of sub-series. Calculations in this case can become very complicated.

In practice the method used is dependent on the situation as different methods work better for different hierarchies. In large unbalanced hierarchies the iterative approach is favoured because of the over complication of the ANOVA approach. But for smaller hierarchies the sparse array approach is seen to be sufficient.

Forecast Performance Evaluation

There have been many comparisons of the well known top-down and bottom-up methods which contrast and name one over the other as the better method. Section IV of Grunfeld and Griliches (1960) looks at an this comparison in the field of economics. They believed that since the perfection of micro equations is unlikely, aggregation would lead to 'net gain' and thus less accurate results than they would have from a macro equation. This indicates a situation where top-down out performs bottom-up although they do not claim that a perfectly specified micro system would not out perform a macro equation. On the other hand Edwards and Orcutt (1969) argue that the bottom-up method is more accurate because of the loss of information in the top-down method. The work from Hyndman et al. (2011) agrees with this.

The data set Hyndman et al. (2011) discussed was from Australian domestic tourism. It concentrates on quarterly domestic tourism demand, measured by the number of visitor nights Australians spend away from their home state. To compare the three methods mean absolute percentage error (MAPE) was used, where smaller MAPE illustrates a better forecast. In this example it was found that the top-down method performed significantly worse than the bottom-up and optimal combination approaches. The only level top-down performed best at was the top level, level 0. For the other levels both the bottom-up and optimal combination were better. It is thought that bottom-up performed particularly well in this example because the data had strong seasonality and trend, people tend to choose where to travel depending on the season and popularity. Hyndman et al. (2011) indicate that with more noisy data optimal combination should be the clear front runner. Prediction intervals could also be calculated when considering optimal combination, giving some idea of the accuracy of the forecast. This would not be easy with the previous two methods.

Conclusion

In this paper forecasting within hierarchical models has been discussed focusing on the bottom-up, top-down and optimal combination methods. It can be seen that the best method to use when looking at hierarchical models depends largely on the situation at hand, all three methods have been seen to give good results and out perform the others at different times. It would therefore be advisable to judge each hierarchical forecasting situation differently and choose an appropriate method depending on the properties of the data. When the bottom level data is noisy, top-down may perform best, on the other hand if there is strong seasonality in the data, bottom-up could be a better choice. Optimal combination is designed to perform well in all situations but it is a comparatively more complex method. In some situations it may not be needed if it is suspected another method will perform just as well. It could also be beneficial to use more than one method and compare resultant forecasts.

References

Athanasopoulos, G., Ahmed, R. A., and Hyndman, R. J. (2009). Hierarchical forecasts for australian domestic tourism. *International Journal of Forecasting*, 25(1):146–166.

Edwards, J. B. and Orcutt, G. H. (1969). Should aggregation prior to estimation be the rule? *The Review of Economics and Statistics*, pages 409–420.

Grunfeld, Y. and Griliches, Z. (1960). Is aggregation necessarily bad? *The Review of Economics and Statistics*, pages 1–13.

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. and Athanasopoulos, G. (2014). *Forecasting: principles and practice*. OTexts.

Paige, C. C. and Saunders, M. A. (1982). Algorithm 583: Lsqr: Sparse linear equations and least squares problems. *ACM Transactions on Mathematical Software (TOMS)*, 8(2):195–209.