Making Regional Forecasts Add Up

Tim van Erven Jairo Cugliari
tim@timvanerven.nl Jairo.Cugliari@inria.fr
INRIA team SELECT and Université Paris-Sud

February 28, 2013

Abstract

Suppose we want to predict both the electricity consumption for K regions individually and the total consumption for all the regions together. For operational reasons, it is sometimes required that the prediction for the total consumption is equal to the sum of the predictions for the individual regions, but this constraint might prevent us from making the best possible predictions. In this work, we therefore introduce a new way to work around this restriction, for the case of quadratic loss. The idea is to adjust the regional predictions slightly so that their sum is closer to the prediction we would like to make for the total consumption. The amount of adjustment is determined in a game-theoretically optimal way.

1 Introduction

Consider the problem of predicting the electricity consumption for K regions, and also the total consumption of all the regions together. If our prediction for region $k \in \{1, \ldots, K\}$ is $\hat{y}_k$ and the true consumption is y_k, then our loss for that region will be measured 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 determines the relative importance of the region. Similarly, if we predict $\hat{y}_*$ for the total consumption $y_* = \sum_{k=1}^K y_k$, then our loss is

$$\ell_*(y_*, \hat{y}_*) = a_*(y_* - \hat{y}_*)^2,$$

again with $a_* > 0$. Thus, if $y = (y_1, \ldots, y_k)$ and $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_k, \hat{y}_*)$, then our total loss is

$$\ell(y, \hat{y}) = \sum_{k=1}^K \ell_k(y_k, \hat{y}_k) + \ell_*\big(\sum_{k=1}^K y_k, \hat{y}_*\big). \tag{1}$$

We will assume that *ideal* predictions $\bar{y} = (\bar{y}_1, \ldots, \bar{y}_K, \bar{y}_*)$ are given, where $\bar{y}_k$ is the ideal prediction for region k, and $\bar{y}_*$ is the ideal prediction for the total consumption. The problem we aim to solve is that in general $\sum_k \bar{y}_k$ may be different from $\bar{y}_*$, but, when we make our actual predictions $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_K, \hat{y}_*)$, we have to satisfy the operational constraint

$$\hat{y}_* = \sum_{k=1}^K \hat{y}_k. \tag{2}$$

This is a common constraint in the industry; for example, it was imposed in the Global Energy Forecasting Competition 2012 on Kaggle.com.

Thus we need to map the ideal predictions $\bar{y}$ to real regional predictions $\hat{y}$ that satisfy the constraint (2). In Section 4 we will argue that this is not a trivial problem, because it is indeed realistic that our best prediction for the total $\bar{y}_*$ will be different from $\sum_k \bar{y}_k$.

Related Work Our problem represents a special case of hierarchical time series (HTS) forecasting, as discussed by Hyndman et al. [2011]. But because we will analyse the problem in a game-theoretic way (instead of making distributional assumptions about the true consumptions y), we arrive at a different solution (see Section 2). We clarify the link between the predictions of the two methods in Section 3, and finally, in Section 4, we compare their performance on simulated data.

2 Game-Theoretically Optimal Predictions

We take the ideal predictions $\bar{y}$ to be our gold standard, so ideally we would like to get loss $\ell(y, \bar{y})$, but in reality we will get loss $\ell(y, \hat{y}) = \ell(y, (\hat{y}_1, \ldots, \hat{y}_K, \sum_k \hat{y}_k))$. (Here we have identified $\hat{y}$ with $(\hat{y}_1, \ldots, \hat{y}_K)$, which is possible because of the constraint (2).) We propose to make the difference between $\ell(y, \bar{y})$ and $\ell(y, \hat{y})$ as small as possible when y is chosen in an adversarial way. That is, we choose $\hat{y}$ to achieve

$$V = \inf_{\hat{y}} \sup_{y} \Big\{ \ell\big(y, (\hat{y}_1, \ldots, \hat{y}_k, \sum_k \hat{y}_k)\big) - \ell(y, \bar{y}) \Big\}. \tag{3}$$

The link with game theory is that this is the optimal strategy in a zero-sum game, in which we move first by choosing $\hat{y}$ and then our opponent responds by choosing y.

If we allow y to be unbounded, then $V = \infty$, so this strategy is of no use to us. Luckily, however, in practice it seems reasonable to assume that our ideal predictions $\bar{y}$ will not be too bad, and that the true consumptions y will fall inside the *confidence bands*

$$y_k \in [\bar{y}_k - B_k, \bar{y}_k + B_k] \tag{4}$$

for some positive constants $B_1, \ldots, B_K$. Notice that these confidence bands are *symmetric* around the ideal predictions $\bar{y}_1$, …, $\bar{y}_K$.

Example 1. In the special case that $B_1 = \cdots = B_K = B$ and $a_1 = \cdots = a_K = a$, such that all the regions are treated the same way, the solution to (3) can be computed in closed form and is given by

$$\hat{y}_k = \bar{y}_k + \max\Big(-B, \min\Big\{B, \frac{\frac{1}{a}}{\frac{1}{a_*} + K\frac{1}{a}} z\Big\}\Big),$$

where $z = \bar{y}_* - \sum_k \bar{y}_k$. We see that the predictions $\hat{y}_k$ follow a single equation that depends only on the weights of the regions and on z, except that they are truncated to fall inside the confidence bands (4).

In general, no closed-form solution to (3) is available, but with some work it can be shown that the optimal predictions can be determined by solving a certain least squares problem with L_1-regularization, which can be done efficiently using standard software to compute the LASSO [Tibshirani, 1996] (in its unconstrained formulation).

3 Interpretation as Approximate Projection

If the sizes of the confidence bands B_k are sufficiently large, it can be shown that the optimization problem (3) is approximately equal to an optimization problem with solution

$$\hat{y}_k = \bar{y}_k + \frac{\frac{1}{a_k}}{\frac{1}{a_*} + \sum_k \frac{1}{a_k}} z, \tag{5}$$

where $z = \bar{y}_* - \sum_k \bar{y}_k$. (See also Example 1.) It turns out that this choice for $\hat{y}$ can be given another interpretation as well: it is also the solution to the minimization problem

$$\inf_{\hat{y}} \Big\{ \sum_{k=1}^{K} a_k (\hat{y}_k - \bar{y}_k)^2 + a_* \big(\sum_{k=1}^{K} \hat{y}_k - \bar{y}_*\big)^2 \Big\},$$

which can be interpreted as an L_2-projection of $\bar{y}$ unto the hyperplane specified by the constraint (2) that takes the regional weights a_k and a_* into account.

In our context, the HTS solution proposed by Hyndman et al. [2011] reduces to $\hat{y}_k = \bar{y}_k + \frac{1}{K+1} z$. Comparison with (5) reveals that this corresponds to the same L_2-projection, except that it assumes all the regional weights to be equal.

4 Experiment with Simulated Data

Data To compare the performance of our method with HTS, we simulate data for two regions:

$$y_1 = 1 + 5x + \sigma\xi + \tau\zeta_1 \qquad\qquad y_2 = 1 + 5x - \sigma\xi + \tau\zeta_2,$$

where ξ, ζ_1 and ζ_2 are independent random variables that are uniformly distributed on $[-1, 1]$ and σ and τ are nonnegative scale parameters. Notice that the noise that depends on ξ cancels from the total consumption $y_1 + y_2$, which makes the total consumption easier to predict than the individual regions. We sample a train set of size 100 for the fixed design $x \in \{1/100, 2/100, \ldots, 1\}$ and a test set of the same size for $x \in \{1 + 1/100, \ldots, 2\}$.

Fitting Models on the Train Set To construct our ideal predictions $\bar{y}_1$ and $\bar{y}_2$, we use the LASSO with cross-validation to calibrate the amount of penalization. To estimate the confidence bands B_1 and B_2, we (somewhat crudely) use the maximum absolute value of the residuals in each region.

If we would also use the LASSO directly to predict the total consumption $y_1 + y_2$, it might not do better than simply using the *bottom-up predictor* $\bar{y}_1 + \bar{y}_2$. We can be sure to do better, however, by adding $\bar{y}_1$ and $\bar{y}_2$ as covariates, such that we fit functions of the form

$$\beta_0 + \beta_1 * x + \beta_2 \bar{y}_1 + \beta_3 \bar{y}_2.$$

Moreover, we introduce prior knowledge into the LASSO by regularizing by $|\beta_0| + |\beta_1| + |\beta_2 - 1| + |\beta_3 - 1|$ instead of $|\beta_0| + |\beta_1| + |\beta_2| + |\beta_3|$, which would make it behave like the bottom-up predictor in the absence of any data.

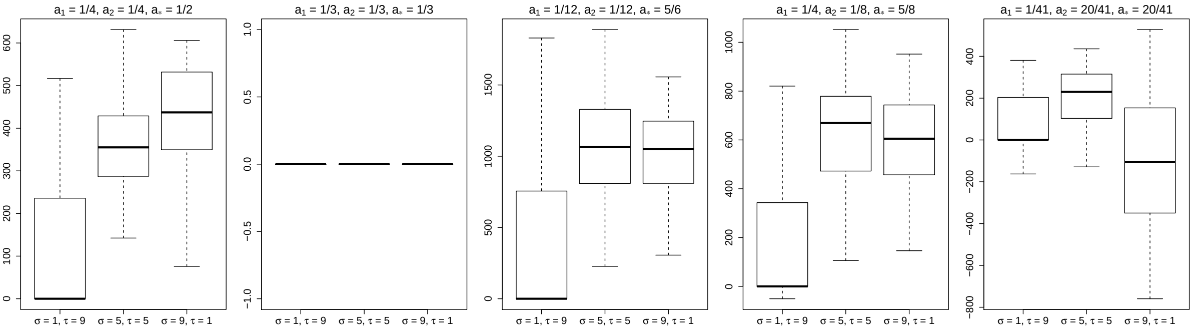

Figure 1: Test loss for HTS minus test loss for our method

Evaluation on the Test Set On the test set, we first use the fitted regression parameters from the Lasso to get ideal predictions $\bar{y}_1$ and $\bar{y}_2$ for the regions. Then we use these regional predictions as covariates to get the ideal prediction $\bar{y}_*$ for the total consumption using the parameters $(\beta_0, \beta_1, \beta_2, \beta_3)$ we found on the train set. We now compute the test set loss for our method (L) and for HTS (L_{hts}) by summing (1) over the test set.

Results It remains to choose the weights a_1, a_2 and a_*, and the scales σ and τ for the noise variables. We repeat the experiment 100 times for different choices of these parameters. The different values for a_1, a_2 and a_* represent different possible electricity network configurations, and are normalized to sum to 1, because scaling them just scales all losses by the same factor. Figure 1 shows box plots that summarize the values of $L_{\text{hts}} - L$ during these 100 repetitions.

We see that our method outperforms HTS in most cases. A notable exception is when $a_1 = a_2 = a_* = 1/3$, for which HTS equals (5), and apparently the bounds B_1 and B_2 are sufficiently large for our method to be the same as well (see the discussion in Section 3). So in this case the methods coincide.

Acknowledgements We would like to thank Mesrob Ohannessian for useful discussions, which led to the projection interpretation in Section 3. This work was supported in part by NWO Rubicon grant 680-50-1112.

References

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.

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