Online hierarchical forecasting for power consumption data

Margaux Brégère, Malo Huard (2021)

Introduction

动机

电力预测，以单独家庭为单位的消费不稳定，难以预测，故关注总量和区域用电量，分层预测。

文献综述

分层预测

- **经典方法**：

  自下而上（bottom-up）：对低级别预测，之后求和 (Dunn，Williams，&Dechaine，1976)；

  自上而下（top-down）：预测聚合级别，之后确定低聚合级别的比例并分解 (Gross & Sohl，1990)；

  其中自下而上的方法可能更好 (Shlifer & Wolff, 1979)；在负荷预测上也被证明其成功性 (Auder, Cugliari, Goude & Poggi, 2018)；

- **最近的分层预测方法**：

  最小迹（MinT）：利用投影进行分层约束的协调（reconcile），通过基本预测的协方差矩阵信息进行 (Wickramasuriya, Athan asopoulos, Hyndman, 2019)，包括正交投影、斜投影（Panagiotelis，Athan asopoulos，Gamakumara，&Hyndman，2020)；

  基于博弈理论的最优协调方法 (Van Erven, Cugliari, 2015)：首先，在不考虑分层约束的情况下，给出时间序列的预测结果；然后利用调和过程使预测结果总体一致。

聚合方法

Vovk(1990)、Cover(1991)、Littlestone和Warmuth(1994)提出了单个序列的聚合方法，不依赖于观测的任何建模即可生成独立于数据生成过程的组合预测。有文献证明其有效性(Mallet，Stoltz&Mauricette，2009; Devaine，Gaillard，Goude，&Stoltz，2013; Gaillard，Goude，&Nedellec，2016)；Goehry，Goude，Massart和Poggi(2020) 将这种聚合方法扩展到分层预测（自下而上）。

本文方法简述

分为三个阶段：

1. 用广义加性模型生成每层的基本预测；
2. 对每个序列使用聚合方法以找到最优线性组合（本文创新部分），使用 ML-poly算法（多项式加权平均）；
3. 将组合预测投影以保证组合满足分层约束（1和3相当于MinT的最小二乘版本）。

Methodology

层次关系的表示

一个三层的层次结构，可以按两种类别构造出两种分层方式，如下图：

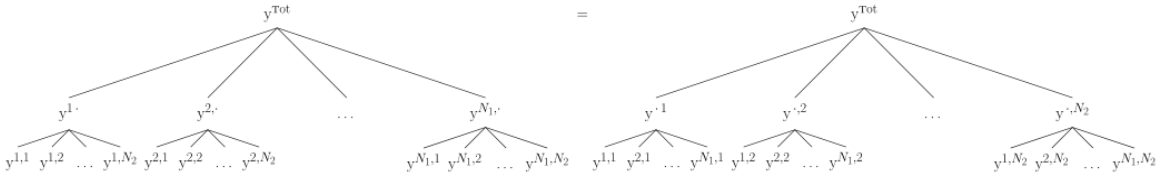


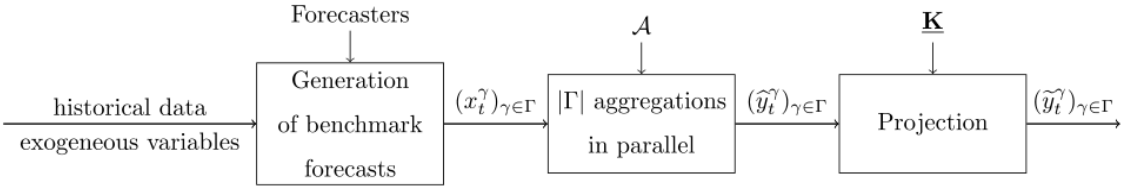
Fig. 2. Representation of two crossed hierarchies.

为表示层次关系，使用矩阵 $\underline{\boldsymbol{K}}$ 与等式约束 $\underline{\boldsymbol{K}}\boldsymbol{y_t} = \mathbf{0_{2+N_1+N_2}}$ ，上图所需的矩阵如下所示。每行代表一个等式约束。

$$
\underline{\mathbf{K}} = \begin{pmatrix}
-1 & \overbrace{1 \cdots 1}^{N_1} & & & & & & & \\
 & -1 & & & \overbrace{1 \cdots 1}^{N_2} & & & & \\
 & & \ddots & & & & \ddots & & \overbrace{1 \cdots 1}^{N_2} \\
 & & & -1 & & & & & \\
-1 & & & & \overbrace{1 \cdots 1}^{N_2} & & & & \\
 & & & & -1 & & 1 & & 1 \\
 & & & & & \ddots & & \cdots & & \ddots \\
 & & & & & & -1 & & 1 & & 1
\end{pmatrix}
$$

三阶段预测方法

方法流程图如下：



- 产生基准预测：利用某一结点的历史数据与外生变量，利用广义加性模型得到基准预测，将每个节点的预测结果组合成向量来进入下一步的聚合预测。$\boldsymbol{x}_t = (x_t^\gamma)_{\gamma\in\Gamma}$ 。
- 聚合预测：考虑到不同节点的相关性与受约束关联的情况，故采用基准的线性组合形式。每个节点的聚合预测定义为$\hat{y}_t^\gamma = \hat{\boldsymbol{u}}_t^\gamma \boldsymbol{x}_t$ ，其中 $\hat{\boldsymbol{u}}_t^\gamma$ 是权重。这一步的结果用向量表示为 $\hat{\boldsymbol{y}}_t = (y_t^\gamma)_{\gamma\in\Gamma}$ 。
- 投影：为保证分层约束，考虑 $\hat{\boldsymbol{y}}_t$ 到 $\underline{\boldsymbol{K}}$ 的正交投影为 $\underline{\boldsymbol{K}}^T(\underline{\boldsymbol{K}}\underline{\boldsymbol{K}}^{\boldsymbol{T}})^{-1}\underline{\boldsymbol{K}}$ ，则到其核的正交投影为 $\boldsymbol{\Pi}_{\underline{\boldsymbol{K}}}(\hat{\boldsymbol{y}}_t) = \boldsymbol{I} - \underline{\boldsymbol{K}}^T(\underline{\boldsymbol{K}}\underline{\boldsymbol{K}}^{\boldsymbol{T}})^{-1}\underline{\boldsymbol{K}}$；而最终的预测结果定义为 $\tilde{\boldsymbol{y}}_t = \boldsymbol{\Pi}_{\underline{\boldsymbol{K}}}(\hat{\boldsymbol{y}}_t)$ 。

预测结果评估

最终预测结果的平均损失定义为：

$$\widetilde{L}_T \overset{\text{def}}{=} \frac{1}{T}\sum_{t=1}^{T}\frac{1}{|\Gamma|}\sum_{\gamma\in\Gamma}\left(y_t^\gamma - \widetilde{y}_t^\gamma\right)^2$$

自然的想法是将上式与其它预测结果比较，如基准预测结果，其平均损失为下式。

$$L_T\left(\left(\delta^\gamma\right)_{\gamma\in\Gamma}\right) \overset{\text{def}}{=} \frac{1}{T}\sum_{t=1}^{T}\frac{1}{|\Gamma|}\sum_{\gamma\in\Gamma}\left(y_t^\gamma - x_t^\gamma\right)^2$$

但是基准预测及其它预测不一定满足分层约束，故直接相比有失公平。故考虑一个集合 C ，其包含所有满足分层约束的常数策略，且其中元素可表示为 $\underline{U}$ 。将预测的损失分解为“近似损失”和“序列损失”后，若要保证最终预测结果最优，需要满足以下关系：

$$\widetilde{L}_T \leqslant \inf_{\underline{\mathbf{U}}\in\mathcal{C}}\left\{L_T(\underline{\mathbf{U}}) + \mathcal{E}_T(\underline{\mathbf{U}})\right\}, \quad \text{where} \quad \mathcal{E}_T(\underline{\mathbf{U}}) = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$$

可比集合 C 定义如下，以保证满足分层约束。

$$\mathcal{C} \overset{\text{def}}{=} \left\{\underline{\mathbf{U}} = \left(\mathbf{u}^1 \,\middle|\, \ldots \,\middle|\, \mathbf{u}^{|\Gamma|}\right) \,\middle|\, \text{Im}\left(\underline{\mathbf{U}}^{\mathrm{T}}\right) \subset \text{Ker}\left(\underline{\mathbf{K}}\right)\right\}$$

而优化损失可以等价为是如下“后悔值”尽可能小：

$$\begin{aligned} R_T(\underline{\mathbf{U}}) &\overset{\text{def}}{=} T|\Gamma| \times \left(\widetilde{L}_T - L_T(\underline{\mathbf{U}})\right) \\ &= \sum_{t=1}^{T}\left\|\mathbf{y}_t - \widetilde{\mathbf{y}}_t\right\|^2 - \sum_{t=1}^{T}\left\|\mathbf{y}_t - \underline{\mathbf{U}}^{\mathrm{T}}\mathbf{x}_t\right\|^2 \end{aligned}$$

而此后悔值满足：假设当每个预测节点的后悔值有上界时，预测结果总体后悔值的上界不大于单节点上界之和。以上证明说明**通过遗憾界限可以改进均方根误差。**

Polynomially weighted average forecaster with multiple learning rates (ML-Poly)

在应用聚合算法时，首先进行真实值与基准预测的标准化。该标准化的形式比较特殊，其目的是为了满足“后悔值”计算的理论依据，保证有界性假设的合理性。同时作者认为这种预处理简化了聚合过程中的超参数搜索步骤。

$$\begin{aligned} y_t^\gamma \to \breve{y}_t^\gamma &\overset{\text{def}}{=} \frac{y_t^\gamma - x_t^\gamma}{S^\gamma} && \text{Observations tranform} \\ \mathbf{x}_t \to \breve{\mathbf{x}}_t &\overset{\text{def}}{=} \underline{\breve{\mathbf{E}}}\mathbf{x}_t && \text{Benchmarks transform} \end{aligned}$$

$$\text{with} \quad S^\gamma = \max_{1-T_0\leqslant t\leqslant 0}|y_t^\gamma - x_t^\gamma| \qquad \text{and} \qquad \underline{\breve{\mathbf{E}}} \overset{\text{def}}{=} \left(\frac{1}{T_0}\sum_{t=1-T_0}^{0}\mathbf{x}_t\,\mathbf{x}_t^{\mathrm{T}}\right)^{-1/2}$$

赋权的算法如下，利用原始数据、基准预测、聚合预测的迭代产生权重。（这里的聚合指的是层次结构中一个节点可表示为结构中所有节点基本预测的线性组合，不是指一个节点产生多个预测）。

Algorithm 1 Polynomially weighted average forecaster with Multiple Learning rates and gradient trick

aim: Predict the time series $\left(y_t^\gamma\right)_{1\leqslant t\leqslant T}$

parameter: Bound E

initialization

$\mathbf{u}_1^\gamma = (1/|\Gamma|, \ldots, 1/|\Gamma|)$

$\widehat{y}_1^\gamma = \mathbf{u}_1^\gamma \cdot \mathbf{x}_1$

$\forall\, i \in \Gamma,\ \widetilde{R}_0^{\gamma,i} = 0$ and $\eta_0^{\gamma,i} = 0$

for $t = 1, \ldots, T-1$ **do**

For each $i \in \Gamma$, update the cumulative regret of benchmark i

$$\widetilde{R}_t^{\gamma,i} = \widetilde{R}_{t-1}^{\gamma,i} + \widetilde{r}_t^{\gamma,i} \quad \text{where} \quad \widetilde{r}_t^{\gamma,i} \stackrel{\text{def}}{=} 2(\widehat{y}_t^\gamma - y_t^\gamma)(\widehat{y}_t^\gamma - x_t^i)$$

For each $i \in \Gamma$, compute the learning rate $\eta_t^{\gamma,i} = \left(E + \sum_{s=1}^t \left(\widetilde{r}_t^{\gamma,i}\right)^2\right)^{-1}$

Compute the weight vector $\mathbf{u}_{t+1}^\gamma = (u_{t+1}^{\gamma,i})_{i\in\Gamma}$ defined as

$$u_{t+1}^{\gamma,i} = \frac{\eta_t^{\gamma,i}\left(\widetilde{R}_t^{\gamma,i}\right)_+}{\sum_{j\in\Gamma} \eta_t^{\gamma,j}\left(\widetilde{R}_t^{\gamma,j}\right)_+}$$

Output prediction $\widehat{y}_{t+1}^\gamma = \mathbf{u}_{t+1}^\gamma \cdot \mathbf{x}_{t+1} = \sum_{i\in\Gamma} u_{t+1}^{\gamma,i} x_{t+1}^i$

Experiments

数据集与实验细节

数据概况：英国智能电表数据，1600个家庭，2009.4.20-2010.7.31数据，半小时记录一次。

包含地区、温度（原始与指数平滑版本）、能见度、湿度、日期、一天的第几个半小时、周几、该日在当年的相对位置等外生变量。

已经有地区可以作为分层的一个类；为了构造另一个类，使用电力消费数据进行矩阵分解提取特征——K-means聚类的方法构造家庭用电行为类。至此可以建立三层结构。

训练集为2009.4.20-2010.4.19，测试集中2010.4.20-4.30用于初始化，报告结果以2010.5.1-2010.7.31为准。

如果某个节点涉及来自不同地区的气象变量，按照电力消费的历史比例加权处理。

基准模型考虑电力消费、气象、日期变量；考虑现实得到电力数据有滞后，故采用滞后48期（一天）的数据进行预测，但每次预测为半小时后数据。

基准预测按照前述方法标准化，之后利用搜索的方式优化聚合算法超参。

实验结果

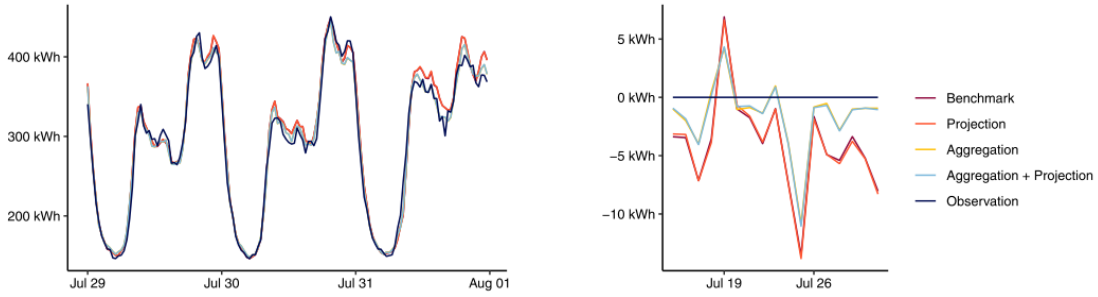
方法对比

首先将四种方法得到的预测结果进行对比：只进行基准预测、基准预测后直接投影（相当于MinT的OLS情况）、基准预测后只聚合、以及完整进行本文的三阶段预测。结果显示，对于全部节点、全局预测和单节点预测结果，**投影都可以符合理论地改进预测**（无论是否聚合）；此外，**三阶段预测确实既能满足分层约束，又能减少误差。**

Table 3
$E_T(\Gamma) \pm \sigma_T(\Gamma)/\sqrt{T}$ (left – see Eq. (24)), $E_T(\{\mathcal{I}\}) \pm \sigma_T(\{\mathcal{I}\})/\sqrt{T}$ (middle), $E_T(\Gamma_0) \pm \sigma_T(\Gamma_0)/\sqrt{T}$ (right) (see Eq. (24)) for "Region + Behavior" clustering for the four strategies defined in Section 5.4 ("Benchmark", "Projection", "Agregation" and "Aggregation + Projection"). $E_T(\Gamma)$ corresponds to $\widetilde{L}_T \times |\Gamma|$ for the "Aggregation + Projection" strategy. The prediction error $E_T(\{\mathcal{I}\})$ corresponds to the mean squared error (over the testing period) of the global consumption and $E_T(\Gamma_0)$ corresponds to a prediction error associated with local consumption forecasts. The dark gray area corresponds to the best prediction error of the column.

	$\mathrm{E}_\mathrm{T}(\Gamma) \pm \frac{\sigma_T(\Gamma)}{\sqrt{T}}$	$\mathrm{E}_\mathrm{T}(\{\mathcal{I}\}) \pm \frac{\sigma_T(\{\mathcal{I}\})}{\sqrt{T}}$	$\mathrm{E}_\mathrm{T}(\Gamma_0) \pm \frac{\sigma_T(\Gamma_0)}{\sqrt{T}}$
Benchmark	455.5 ± 1.1	205.8 ± 9.3	66.3 ± 0.1
Projection	450.7 ± 1.1	200.8 ± 9.2	66.3 ± 0.1
Aggregation	397.9 ± 1.0	172.0 ± 8.6	61.2 ± 0.1
Aggregation + Projection	396.0 ± 1.0	170.3 ± 8.5	61.1 ± 0.1

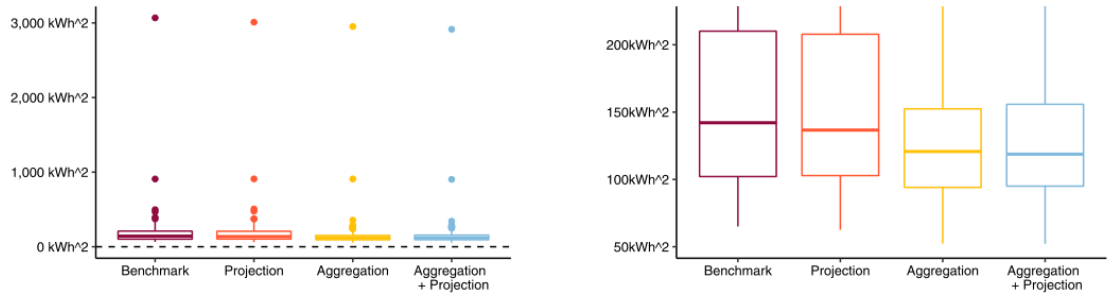
观察预测结果的对比图，可以发现，聚合对预测影响改进更大，投影次之。此外，聚合还有减小预测方差的好处。



(a) Forecasts at half-hour intervals on three days.

(b) Daily average signed errors on a week.

Fig. 3. Forecasts and errors associated with the four strategies "Benchmark", "Projection", "Aggregation", and "Aggregation + Projection" and observations of global consumption ($\gamma = \mathcal{I}$) at the end of the test period.



(a) Original boxplots.

(b) Boxplots trimmed at 220 kWh^2.

Fig. 4. Distribution over the test period of the daily mean squared error of global consumption for the four strategies "Benchmark", "Projection", "Aggregation", and "Aggregation + Projection".

之后是对于不同分层结构（区域，行为，区域与行为三种结构）的改进结果比较。这里只比对全局预测的结果。对于自下而上的预测方法，仅考虑区域信息最好；而其它方法建议使用区域与行为的综合信息。

Table 4
$E_T(\{\mathcal{I}\}) \pm \sigma_T(\{\mathcal{I}\})/\sqrt{T}$ (see Eq. (24)) for the fives strategies defined in Section 5.4 ("Benchmark", "Bottom-up, "Projection", "Agregation" and "Aggregation + Projection"), with benchmark predictions ($x_t^{[\mathcal{I}]}$ that are the same for all clusterings) made with General Additive Models and aggregated with ML-Pol algorithm, for the three segmentations ("Region", "Behavior" and "Region + Behavior"). The prediction error $E_T(\{\mathcal{I}\})$ corresponds to the mean squared error (over the testing period) of the global consumption. The dark gray area corresponds to the best prediction error of the table and the light gray area to the best one, for a given strategy.

Clustering	Benchmark	Bottom-up	Projection	Aggregation	Aggregation + Projection
Region	205.8 ± 9.3	189.9 ± 8.3	201.3 ± 9.1	187.8 ± 8.4	186.7 ± 8.4
Behavior	–	208.4 ± 9.6	205.2 ± 9.3	179.3 ± 8.4	179.3 ± 8.4
Region + Behavior	–	$\mathbf{201.0 \pm 8.5}$	$\mathbf{200.8 \pm 9.2}$	$\mathbf{172.0 \pm 8.6}$	$\mathbf{170.3 \pm 8.5}$