Adaptive Conformal Inference for Multi-Step Ahead Time-Series Forecasting Online

Johan Hallberg Szabadváry

Jönköping University, Stockholm University

Introduction

The on-line setting

Reality outputs a sequence of pairs

$$(x_1, y_1), (x_2, y_2), \ldots$$

called *examples*, consisting of *objects* x and *labels* y. For short, set $z_i := (x_i, y_y)$.

Our standard assumption is that examples are drawn from some probability distribution P on the example space Z that is exchangeable.

Adaptive conformal inference (ACI)

ACI was proposed by Gibbs and Candès as a way to achieve asymptotic validity for non-exchangeable data. It uses a simple online update of the significance level ε:

$$\varepsilon_{n+1} = \varepsilon_n + \gamma(\varepsilon - \mathrm{err}_n^{\varepsilon_n}(\Gamma)),$$

where γ is a step size.

For all $N \in \mathbb{N}$,

$$\left| \varepsilon - \frac{1}{N} \sum_{i=1}^{N} \mathrm{err}_n^{\varepsilon_n}(\Gamma) \right| \leq \frac{\max\{\varepsilon_1, 1 - \varepsilon_1\} + \gamma}{\gamma N} \ a.s.$$

In particular, ACI ensures that a conformal predictor is asymptotically valid, even if exchangeability is violated.

Adapting ACI for multi-step ahead time-series forecasting

Multi-step ahead setup

Let $w_1, w_2, \ldots$ be observations of a time-series, and suppose we want to predict h steps ahead, using e.g. p lagged values and/or some exogenous variables.

An object x_t then consists of p lagged values. Its label $y_t = (y_{t,1}, \ldots, y_{t,h})$ is a vector of the next h values.

The problem

We use x_t to predict $y_t = (y_{t,1}, \ldots, y_{t,h})$ at significance levels $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_h)$.

Problem: We only get to observe $y_{t,1}$. No way of knowing it the other predictions are wrong.

But we know the values $y_{t-h+1,1}, \ldots, y_{t,1}$. Suppose we have predictions made at times $t - h + 1, \ldots, t$.

Error observation

Denote by $l_{t,i}$ and $u_{t,i}$ the lower and upper bounds of our prediction intervals, with $i = 1, \ldots, h$. Consider the matrices,

$$L_t = \begin{pmatrix} l_{t,1} & l_{t,2} & \cdots & l_{t,h} \\ l_{t-1,1} & l_{t-1,2} & \cdots & l_{t-1,h} \\ \vdots & \vdots & \ddots & \vdots \\ l_{t-h+1,1} & l_{t-h+1,2} & \cdots & l_{t-h+1,h} \end{pmatrix}$$

and

$$U_t = \begin{pmatrix} u_{t,1} & u_{t,2} & \cdots & u_{t,h} \\ u_{t-1,1} & u_{t-1,2} & \cdots & u_{t-1,h} \\ \vdots & \vdots & \ddots & \vdots \\ u_{t-h+1,1} & u_{t-h+1,2} & \cdots & u_{t-h+1,h} \end{pmatrix}.$$

The diagonal elements are the lower and upper bounds of the predictions made at times $t - h + 1, \ldots, t$ for the value $y_{t,1}$.

Online update of significance levels

Define the vector of errors known at time t as

$$\mathbf{err}_t = (\mathrm{err}_{t,1}^{\varepsilon_{t,1}}, \mathrm{err}_{t-1,2}^{\varepsilon_{t-1,2}}, \ldots, \mathrm{err}_{t-h+1,h}^{\varepsilon_{t-h+1,h}}) = \neg\Big(\mathrm{diag}(L_t) \leq y_{t,1} \leq \mathrm{diag}(U_t)\Big)$$

where "$\neg$" denotes logical negation, and the comparisons are made element-wise.

Now we can update

$$\boldsymbol{\varepsilon}_{t+1} = \boldsymbol{\varepsilon}_t + \boldsymbol{\gamma}(\boldsymbol{\varepsilon} - \mathbf{err}_t)$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_h)$ is a vector of step sizes, and vector multiplication is understood as element-wise multiplication.

Initialisation

For $t = 1, \ldots, h$ we do not have the entire error vector. Note that $\mathbf{err}_t = \boldsymbol{\varepsilon}$ is a fixed point of the iteration.

For any prediction step i where we can not observe the error, set $\mathrm{err}_{t,i} = \varepsilon_i$. This keeps the significance level at prediction step i fixed until the relevant information arrives.

Validity and finite sample guarantees

All we do is apply ACI to each prediction step, using the incoming errors as soon as they arrive. Therefore, we have

$$\left| \frac{1}{T} \sum_{t=i}^{T} \mathrm{err}_{t,i} - \varepsilon_i \right| \leq \frac{\max\{\varepsilon_{i,i}, 1-\varepsilon_{i,i}\}+\gamma_i}{\gamma_i T}, \quad i = 1, \ldots, h \quad (a.s.).$$

An upper bound on the overall error rate is obtained by combining these inequalities

$$\left| \frac{1}{T} \sum_{t=h}^{T} \left(\frac{1}{h} \sum_{i=1}^{h} \mathrm{err}_{t,i} \right) - \frac{1}{h} \sum_{i=1}^{h} \varepsilon_i \right| \leq \frac{1}{h} \sum_{i=1}^{h} \frac{\max\{\varepsilon_{h,i}, 1-\varepsilon_{h,i}\} + \gamma_i}{\gamma_i T} \ (a.s.)$$

Summary

- Allowing for different target error rates lets us balance the trade-off between tight intervals and high confidence.
- Different step sizes lets us balance the trade-off between stability and adaptability that comes with ACI.
- Same finite sample guarantee as ACI on individual prediction steps, which also lets us bound the overall error rate.

The same kind of framework will work for any method that uses ε as control input, e.g. Conformal PID by Angelopolis et. al.

Example: Electricity demand forecasting

Setup

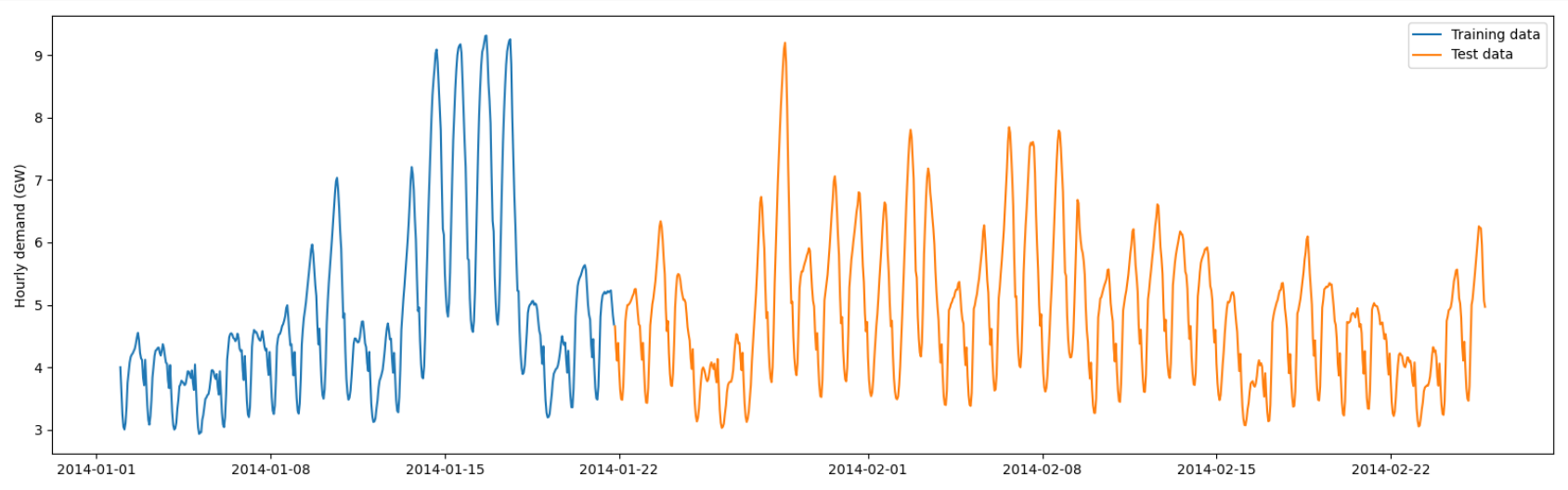

Figure 1: The Victoria electricity demand dataset. We use a MIMO version of the conformailzed ridge regression algorithm (Vovk et. al.). The objects consist of 24 lagged values together with temperature and calendar information. The forecast horizon is five hours.

Example 1. Same target coverage rate and step size for all predictions

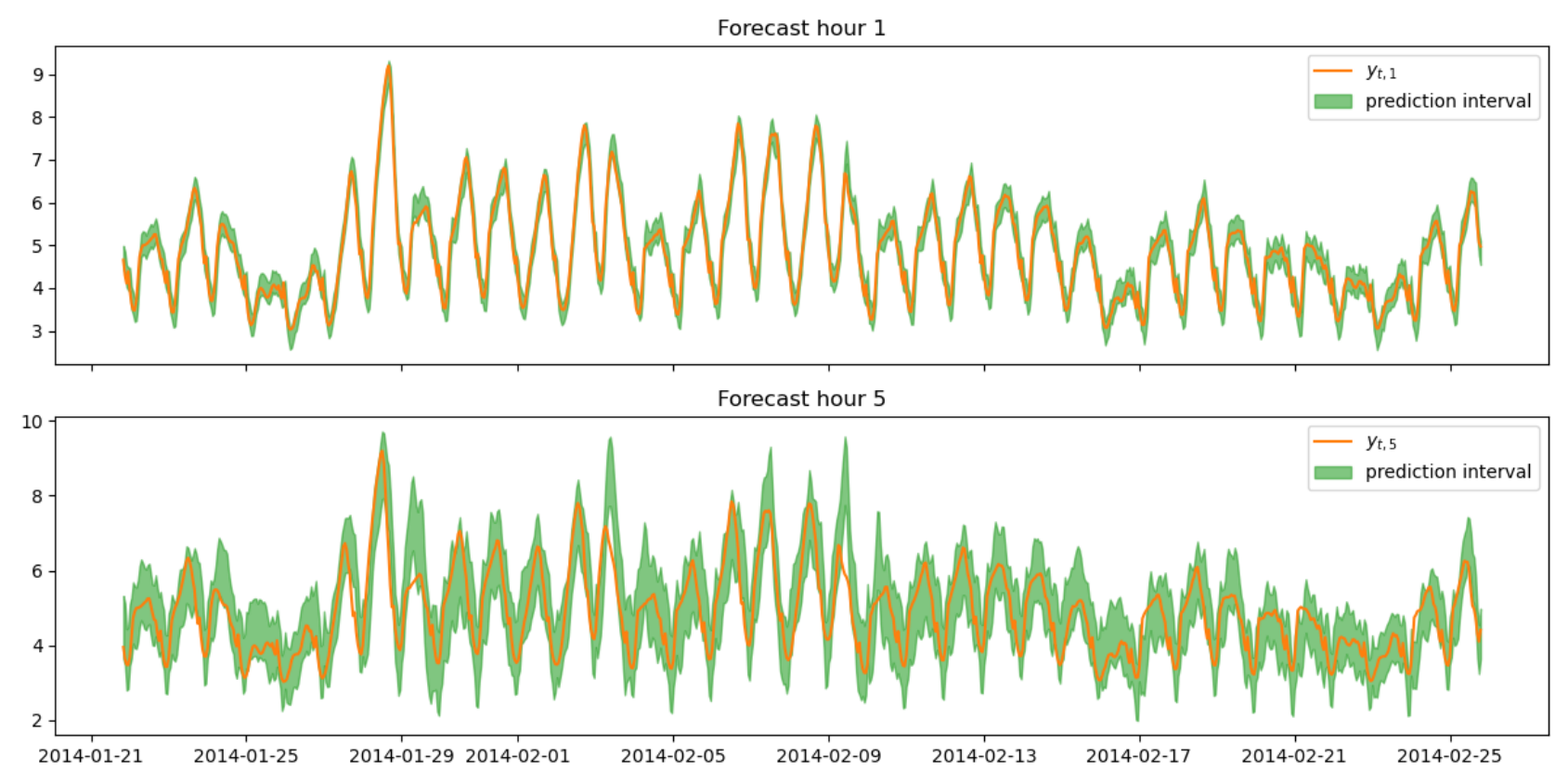

Figure 2: $\varepsilon = (0.1, 0.1, 0.1, 0.1, 0.1)$ and $\gamma = (0.005, 0.005, 0.005, 0.005, 0.005)$

Example 1. Same target coverage rate and step size for all predictions

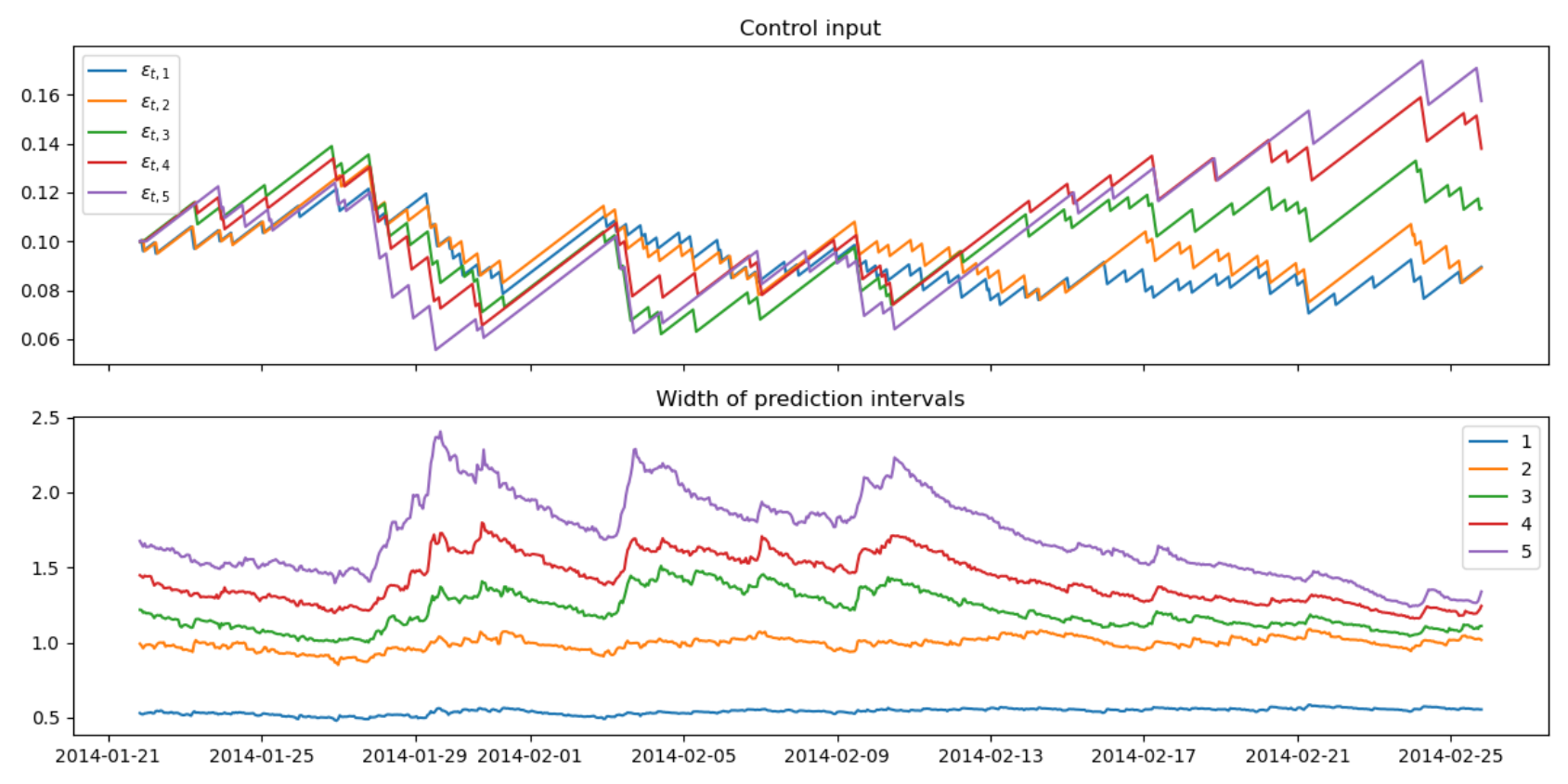

Figure 3: $\varepsilon = (0.1, 0.1, 0.1, 0.1, 0.1)$ and $\gamma = (0.005, 0.005, 0.005, 0.005, 0.005)$

Example 2. Different target coverage but same step size

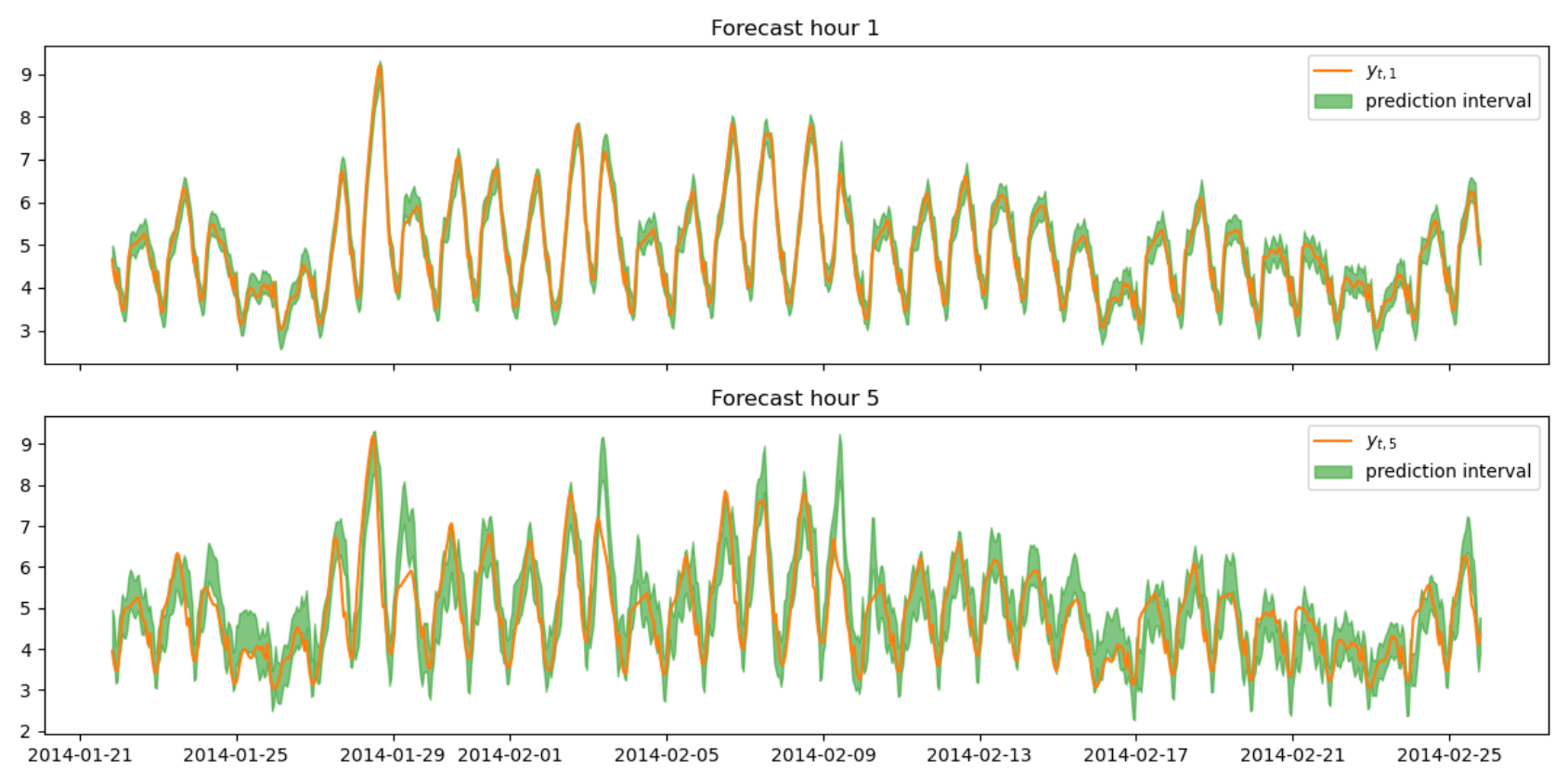

Figure 4: $\varepsilon = (0.1, 0.15, 0.2, 0.25, 0.3)$ and $\gamma = (0.005, 0.005, 0.005, 0.005, 0.005)$

Example 2. Different target coverage but same step size

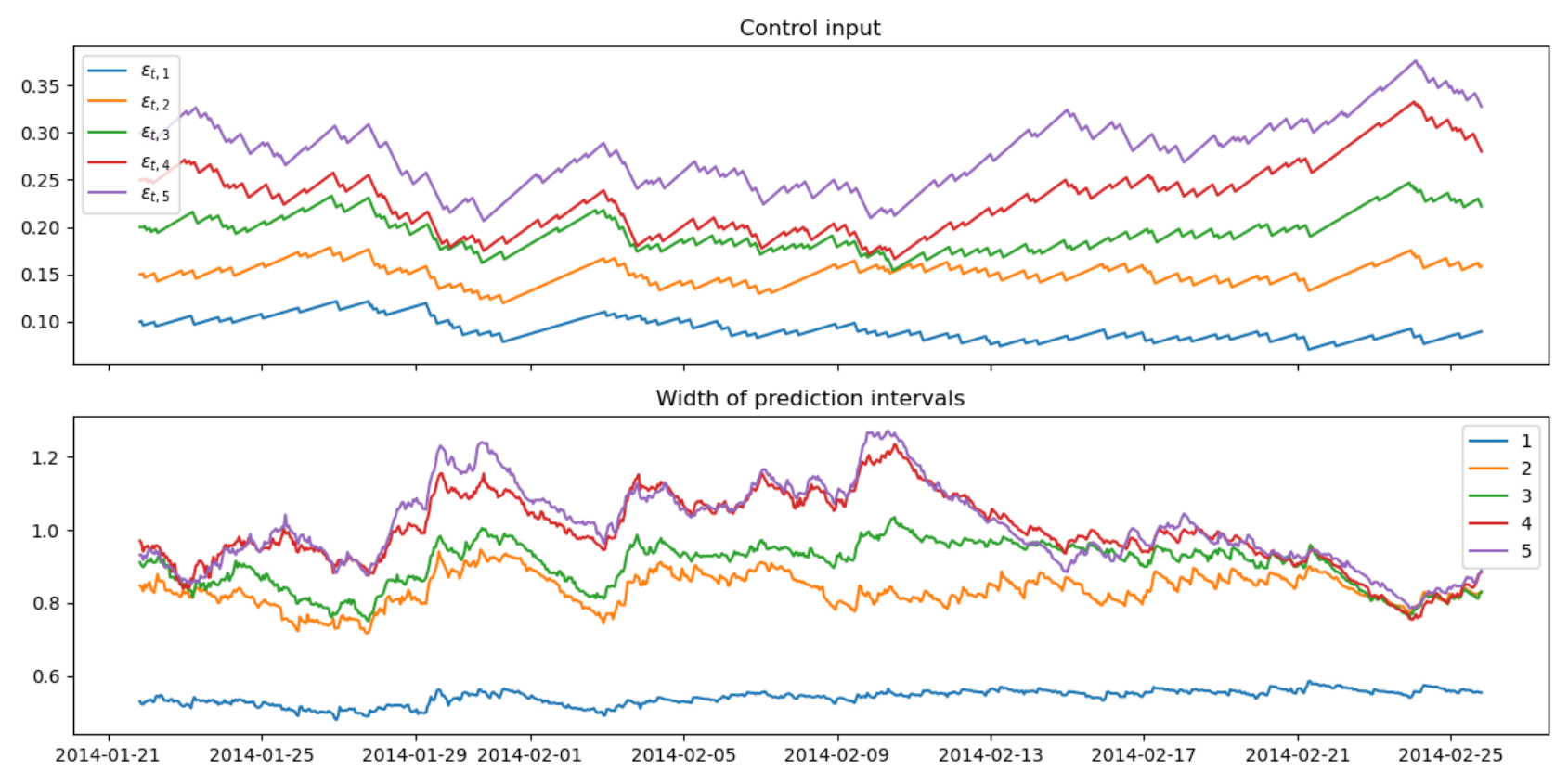

Figure 5: $\varepsilon = (0.1, 0.15, 0.2, 0.25, 0.3)$ and $\gamma = (0.005, 0.005, 0.005, 0.005, 0.005)$

Example 3. Different target coverage and step sizes

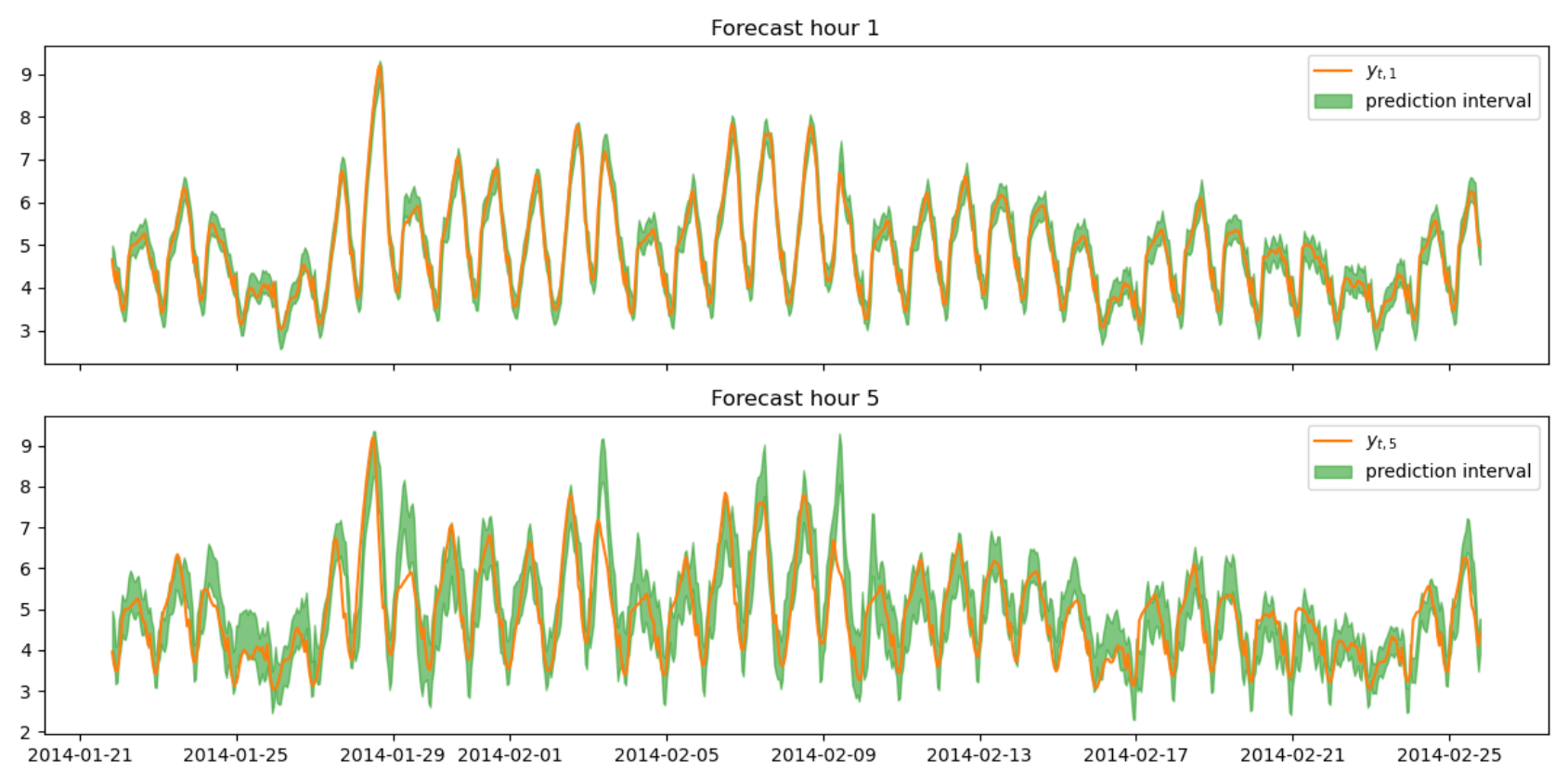

Figure 6: $\varepsilon = (0.1, 0.15, 0.2, 0.25, 0.3)$ and $\gamma = (0.005, 0.007, 0.009, 0.011, 0.013)$

Example 3. Different target coverage and step sizes

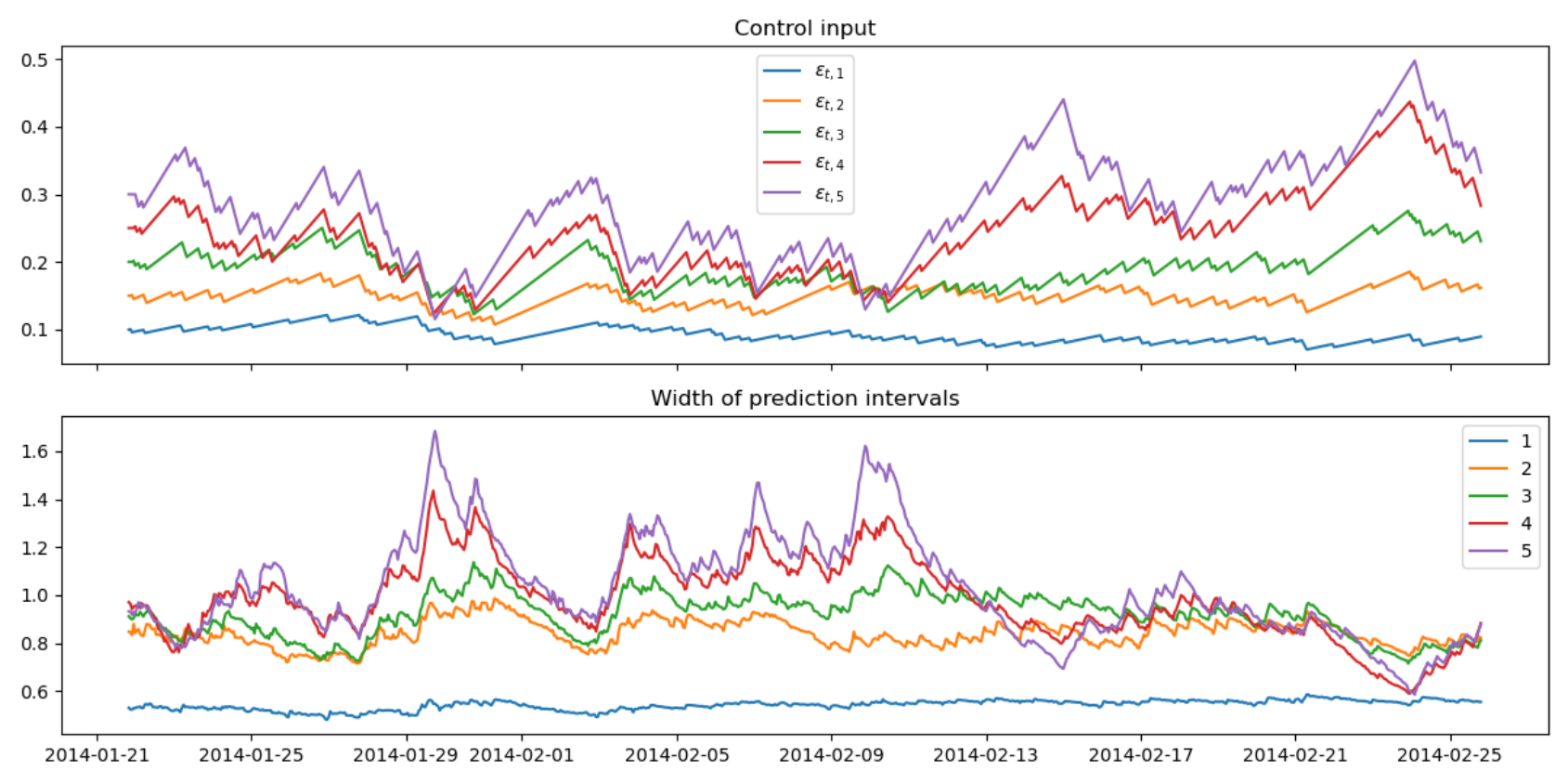

Figure 7: $\varepsilon = (0.1, 0.15, 0.2, 0.25, 0.3)$ and $\gamma = (0.005, 0.007, 0.009, 0.011, 0.013)$

Thank you for your time and attention!