Chapter 1 Experiment

?? showed the number of humans that minimizes the expected squared error of the ensemble in the case where a machine and humans follow the proposed models. However, it is unclear whether the proposed models is appropriate for real problems. Furthermore, even if the proposed model is appropriate, it is impossible to know the true values of parameters $\mathrm{var}(\varepsilon_h)$, $\mathrm{cov}(\varepsilon_h)$ and $\mathrm{cov}(\varepsilon_\theta, \varepsilon_h)$. These can only be *estimated* from past forecasts. The distributions outputted by a machine $f_\theta(y|\boldsymbol{x})$ are also based only on the training set, so they are not always adaptable to unknown inputs. Therefore, we apply the proposed human-machine ensemble method to actual economic forecasts and investigate its performance by comparing with other forecasting methods.

In this chapter, Section 1.1 and Section 1.2 describe data and forecasting methods, respectively. The results are in **??**.

1.1 Data

We used inflation in the U.S. as targets of economic forecasts. This section describes four inflation indicators and two surveys for economic forecasts.

1.1.1 Inflation

We consider four different measures of inflation: CPI-U for All Urban, All Items (*CPI*); CPI-U for All Urban, All items less food and energy (*CoreCPI*); Personal Consumption Expenditures (*PCE*); and PCE excluding food and energy (*CorePCE*). All measures are seasonally adjusted. CPI and CoreCPI are obtained from the Bureau of Labor Statistics[1)] and the sample period is from Jan. 1957 to Oct. 2016. PCE and CorePCE are obtained from the Bureau of Economic Analysis[2)] and the sample period is from Jan. 1959 to Oct. 2016.

CPI measures the total cost of goods and services purchased by urban consumers. It is commonly used as an indicator of inflation. CoreCPI excludes food and energy price from CPI. Short-term fluctuations of CoreCPI is smaller than CPI because the prices of food and energy change easily.

[1)] `https://www.bls.gov/`

[2)] `https://www.bea.gov/`

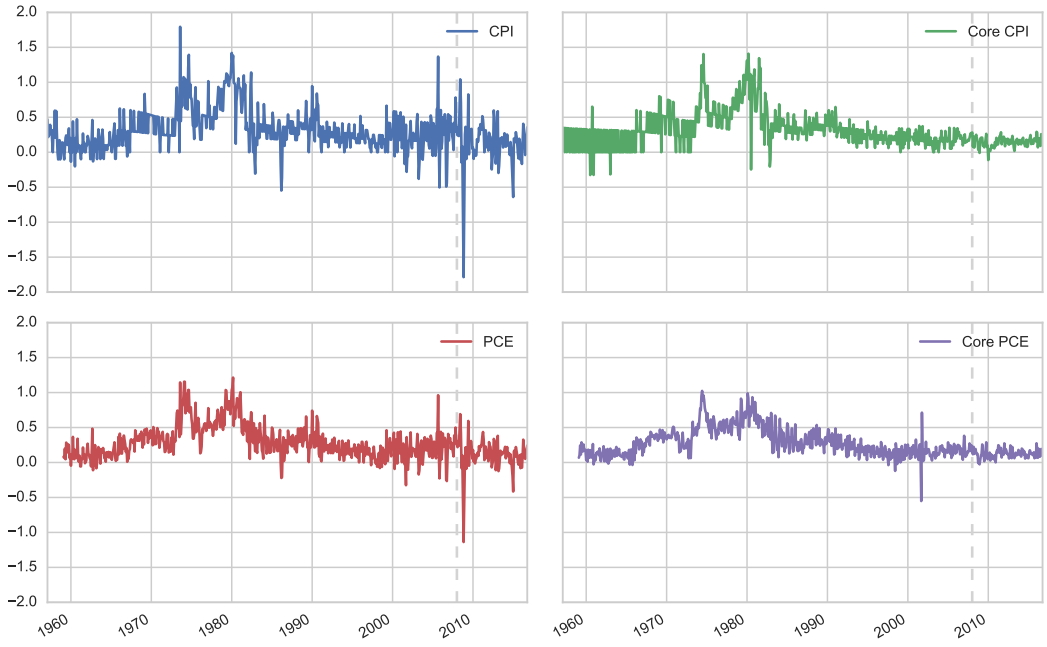

Figure 1.1: Monthly inflation rate measured by CPI, CoreCPI, PCE, and CorePCE. The dotted line represents 2008, which is the boundary between the traing set and test set.

PCE also measures the cost of goods and services purchased by consumers, but includes costs that CPI does not include such as medical expenses, and the data source is different. In addition, calculation procedures are different such that PCE uses chained weights while CPI uses fixed weights. CorePCE exclude food and energy price from PCE.

This thesis defines inflation rate from $t-1$ to t as:

$$\pi_t = \log\left(\frac{P_t}{P_{t-1}}\right) \times 100, \tag{1.1}$$

where P_t is the level of an index at time t. This thesis uses the terms "inflation" and "inflation rate" as defined in (1.1). π_t can be regarded as an approximation of the normal change rate by the following first order Taylor expansion approximation, assuming that the change is sufficiently small,

$$\log\left(\frac{P_t}{P_{t-1}}\right) = \log\left(1 + \frac{P_t - P_{t-1}}{P_{t-1}}\right) \approx \frac{P_t - P_{t-1}}{P_{t-1}}.$$

Figure 1.1 shows monthly inflation of the four indices.

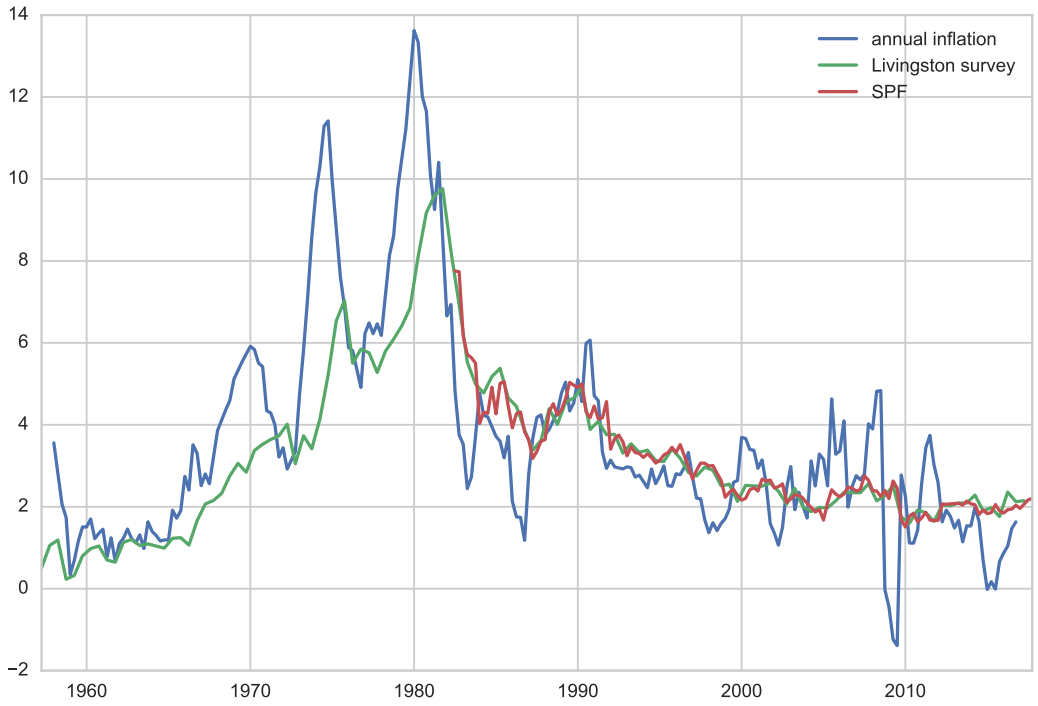

Figure 1.2: Annual CPI change rate and averaged forecasts by the Livingston Survey and the SPF.

1.1.2 Surveys

We used two surveys for economic forecasts: the Livingston Survey[1)] and the Survey of Professional Forecasters[2)] (SPF) as forecasts by humans. The Federal Reserve Bank of Philadelphia has been taking these surveys to economic experts. While the Livingston Survey includes the economists from government and academia, the SPF mainly covers industry. Figure 1.2 shows forecasts by each survey.

Livingston Survey The Livingston Survey is conducted twice a year, in June and in December, from 1946. This study treats from Jun. 1957 to Jun. 2016. The Livingston Survey includes only CPI as a survey item. It was reorganized in 2004, and started to forecast seasonally adjusted series from not seasonally

[1)] `https://www.philadelphiafed.org/research-and-data/real-time-center/livingston-survey`

[2)] `https://www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters/`

adjusted series. It forecasts inflation 6 months later and 12 months later as the short term forecasts. The average number of respondents to 12 months forecast was 45.6 and the standard deviation was 11.6.

SPF The SPF began in 1968 and is conducted quarterly. It covers all four indicators as survey items. However, the survey of CPI started in 1981Q3, and the other three indicators started in 2007Q1. In this thesis, we used all the survey data up to 2015Q4. The average number of respondents to CPI was 34.4 and the standard deviation was 9.3.

1.2 Forecasting Models and Methodology

This section describes four forecasting methods and evaluation method for the experiment. The forecasting methods are a time-series model *ARMA(1,1)* as benchmark, a *recurrent neural networks (RNN)* as machine forecasts, survey forecasts as human forecasts, and the human-machine ensemble method proposed in **??**. We created models that predict annual change rate of the four indices using ARMA(1,1) and RNN, and also created models that predict 6 months change rate of CPI. Due to the restrictions on the survey data, 12 months and 6 months forecasts for CPI are made from the Livingston Survey, and annual forecasts for all indices are made from the SPF. The human-machine ensemble method follows the restrictions of surveys. The data set is divided into the training set and test set, and creation of prediction models does not include the test set. Forecasting accuracy is evaluated by *Root Mean Squared Error (RMSE)* for the test set.

1.2.1 Time-series model

We used an *autoregressive moving average (ARMA) model* for benchmark forecasts as well as Ang et al. [**?**]. ARMA(1,1) is a traditional model for inflation forecasts. Ang et al. made a model based on quarterly inflation rate, and we also did so.

Assume that inflation is stationary to apply an ARMA model. A process is

stationary means that the following equations hold for arbitrary t and k:

$$\begin{aligned} E[\pi_t] &= \mu \\ \mathrm{cov}(\pi_t, \pi_{t-k}) &= E[(y_t - \mu)(y_{t-k} - \mu)] = \lambda_k, \end{aligned} \tag{1.2}$$

where μ and λ_k are constants. This assumption is totally different from the model of **??**, which assumed that the expected error varies depending on the situation t.

The specifications of the ARMA(1,1) model is

$$\pi_{t+1} = \mu + \phi\pi_t + \psi\epsilon_t + \epsilon_{t+1}, \tag{1.3}$$

where ϵ_t is white noise with variance σ^2 [**?**]. The parameters are μ, ϕ, ψ, σ^2, and they are estimated by maximum likelihood conditional on a zero initial residual, assumed the Gaussian white noise.

Applying this model, when an inflation rate at time t, π_t, is given, the forecast value of inflation rate after one period is

$$\hat{\pi}_{t+1|t} = E[\pi_{t+1}|\pi_t] = \mu + \phi\pi_t + \psi\hat{\epsilon}_t. \tag{1.4}$$

$\hat{\epsilon}_t$ is obtained by sequentially approximating like $\hat{\epsilon}_2 = \pi_2 - \mu - \phi\pi_1, \hat{\epsilon_3} = \pi_3 - \mu - \phi\pi_2 - \psi\hat{\epsilon_2}, \ldots$, in which the initial value is $\hat{\epsilon}_0 = 0$. Since the term of ϵ disappears from forecasts after two period, the forecast values can be obtained sequentially by the following relationship:

$$\hat{\pi}_{t+k|t} = \mu + \phi\hat{\pi}_{t+k-1|t}.$$

However, $\hat{\pi}_{t+k}$ is a change rate for one period from $t+k-1$ to $t+k$, so the change rate from t to $t+k$ is the sum of k periods:

$$\hat{\pi}_{t+k,k} = \sum_{i=1}^{k} \hat{\pi}_{t+i}. \tag{1.5}$$

When $k = 2$, it is the 6 months later forecast, and when $k = 4$, it is the 12 months later forecast.

When using ARMA(1,1), the expected squared error of a forecast of one period ahead is always σ^2 regardless of t. Therefore, it is not applicable to the machine model in Section **??**, which assumed the expected error varies depending on the situation. The next section describes RNN that can be used as a machine model for the proposed human-machine ensemble method.

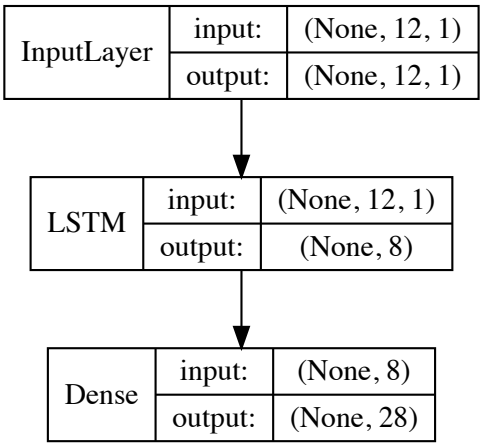

Figure 1.3: The RNN architecture used in the experiment. The inside of parentheses describes the dimentions of data, and the first term represents the batch size 32 instead of "None." The output dimention 28 is for forecasts of annual CPI change rate.

1.2.2 Recurrent neural network model

We used a RNN model that includes *Long short-term memory* (LSTM) units in a hidden layer. Let θ denote the RNN model, which outputs discrete probability distributions $f_\theta(y|\boldsymbol{x})$ when the past 12 months inflation $\boldsymbol{x} = [\pi_{t-11}, \pi_{t-10}, \ldots, \pi_t]$ are given. The forecast value y is an inflation rate for 6 months $\hat{\pi}_{t+6,6}$ or an inflation rate for 12 months $\hat{\pi}_{t+12,12}$. Figure 1.3 shows the RNN architecture that realizes these input and output. By using this architecture, we created five models that forecast annual inflation rates of the four indices, and CPI change rate for 6 months. The rest of this section describes the details of the architecture and how to create the prediction models.

LSTM Ordinary RNN has a problem that, given a long input sequence, the gradient rapidly increases or disappears when backpropagation is applied. It is called *vanishing gradient problem*. The LSTM deals with this problem, and makes long-term memory possible [**?**, **?**]. Replace units in each hidden layer of a basic RNN with LSTM memory blocks shown in Figure 1.4. It looks like there

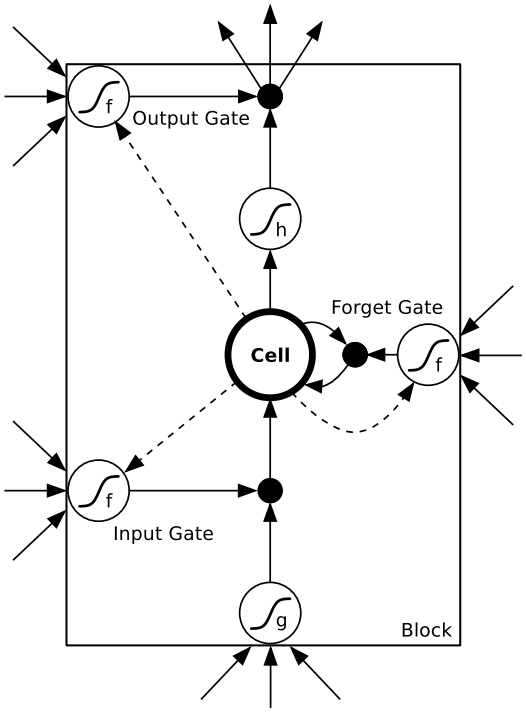

Figure 1.4: LSTM memory block [**?**]

are four inputs, but they are all the same, and identical to the input of a unit in an ordinary RNN's hidden layer. In the center of the block is a cell. Small black circles around the cell are gates, which have activation function f. As logistic sigmoid is usually used for f, the gate activations are between 0 (gate close) and 1 (gate open). The cell stores a state, and realizes the memory by inputting the state to itself after one time interval. However, a forget gate is on the way, and if the activation is close to 0, the state will be forgotten. The outputs of activation functions g and h at the cell input and output are also transmitted if the activation of the gate is close to 1, and are blocked if it is close to 0. With the above mechanism, the LSTM overcomes the limitation of RNN. The experiment used eight LSTM units in the hidden layer, with tanh for the activation functions g and h.

Output layer As mentioned in Section **??**, the activation function of the output layer uses the *softmax function* so that the prediction model can output probability distributions. Let u_k denote input of unit k, and let o_k denote

Table 1.1: The minimum value, maximum value, number of output labels and difference between labels in each index.

	CPI(12M)	CPI(6M)	CoreCPI	PCE	CorePCE
min	0.1	-0.4	0.7	0.5	0.9
max	13.6	7.4	12.7	11.0	9.9
size	28.0	27.0	25.0	22.0	19.0
step	0.5	0.3	0.5	0.5	0.5

output. Then, the softmax function is

$$o_k = \frac{\exp(u_k)}{\sum_{j=1}^{K} \exp(u_j)}, \tag{1.6}$$

where K is the number of units in the output layer. The sum of outputs $v_1, \ldots, v_K$ is always 1. And the output o_k can be interpreted as the probability of belonging to the corresponding class. Making the labels $y_1, \ldots, y_K$ correspond to each unit of the output layer realizes the output of discrete probability distributions. For the labels, separate every 0.5 from the minimum value to the maximum value in the target outputs of training set. For example, when forecasting the annual CPI change rate, since the minimum value and the maximum value in the training set are 0.13 and 13.62 respectively, the values of labels are $[0.1, 0.6, 1.1, \ldots, 13.6]$, and the number of units is $K = 28$. Table 1.1 shows the minimum and maximum values, the number of labels, and the difference between labels in each series. Only when forecasting the 6 months CPI change rate, we set the step as 0.3 to make the dimensions of the outputs similar to others.

Training We divided the data set into the training set and test set. Creation of models uses only training set. The test set is for evaluation. The pairs of input and output are prepared as follows: First, the input is a sequence of monthly inflation for the past 12 months $\boldsymbol{x}_t = [\pi_{t-11}, \ldots, \pi_t]$ from a time t. And these are normalized with the maximum value in the training set as 1. On the other hand, the output is a vector that has K elements $\boldsymbol{d}_t = [0, 0, \ldots, 1, \ldots, 0]$, where the kth value is 1 and the others are 0 because y_k is closest to the correct

value. Furthermore, we used the *cross entropy* as the error function:

$$E(\boldsymbol{w}) = -\sum_{t=1}^{T}\sum_{k=1}^{K} d_{tk} \log v_k(\boldsymbol{x}_t; \boldsymbol{w}). \tag{1.7}$$

The cross entropy is obtained by inverting the log-likelihood of parameters $\boldsymbol{w}$ for the training set $\{(\boldsymbol{x}_t, \boldsymbol{d}_t)\}(t = 1, \ldots, T)$, when assumed the output is posterior distribution. We used the *RMSprop* for the learning algorithm, and set the batch size 32. We stopped learning after 400 epochs, which number is obtained by cross validation.

1.2.3 Survey forecasts

In the Livingston Survey and the SPF, the number of respondents is different for every survey. The minimum number of respondents through the both surveys is 9. Armstrong [**?**] has proposed the principles for combining forecasts, and one of them is "use at least five forecasts when possible." Hence, we sampled five forecasters randomly from each survey to make the number of forecasters equal. The average values are regarded as the forecasts of the survey.

In addition, it is necessary to convert the surveyed values to appropriate inflation rate. The rest of this section describes the methods to calculate the forecast values based on the documentation of each survey.

Livingston Survey The participants in the Livingston Survey respond with not inflation rate but CPI level. The Livingston Survey presents the index level 2 months before to the participants. Based on that level, the participants forecast the value of the current month, $\hat{P}_t$, the one of 6 months after, $\hat{P}_{t+6}$, and the one of 12 months after, $\hat{P}_{t+12}$. However, it was in 1992 that the survey began to include the forecasts of the current month. According to the presence or absence of the forecast value of the current month, the forecast value of m months after, $\hat{\pi}_{t+m,m}$, is calculated by

$$\hat{\pi}_{t+m,m} = \begin{cases} \log\left(\frac{\hat{P}_{t+m}}{\hat{P}_t}\right) & (\text{if } \hat{P}_t \text{ exists}) \\ \log\left(\frac{\hat{P}_{t+m}}{P_{t-2}}\right)^{\frac{m}{m+2}} & (\text{otherwise}).9 \end{cases} \tag{1.8}$$

SPF The participants in the SPF respond with inflation rate, but this is a simple change rate $P_{t+12}/P_t - 1$, not based on the definition of equation (1.1). It

is impossible to convert this value to the inflation rate of (1.1), so the experiment regards it as the approximation. Note that since the forecast values are biased somewhat higher by not taking logarithm, the error becomes slightly larger.

1.2.4 Human-machine ensemble

The Human-machine ensemble method combines the RNN models and individual forecasts from each survey. There are the five RNN models, which forecast annual inflation rate of CPI after 6 months and CPI, CoreCPI, PCE, and CorePCE after 12 months. The Livingston Survey forecasts the index level of CPI after 6 months and 12 months, and the SPF forecasts annual inflation of the four indices.

Execution of the human-machine ensemble method requires the three parameters, $\mathrm{var}(\varepsilon_h)$, $\mathrm{cov}(\varepsilon_h)$ and $\mathrm{cov}(\varepsilon_\theta, \varepsilon_h)$. We estimated these parameters from the forecasts for the training set. Let N denote the number of participants in the training set, and let ε_{h_i} and $\mathrm{var}(\varepsilon_{h_i})$ denote the error and the variance of a participant i respectively. Then, $\mathrm{var}(\varepsilon_h)$ is estimated by

$$\mathrm{var}(\varepsilon_h) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{var}(\varepsilon_{h_i}). \tag{1.9}$$

Also, $\mathrm{cov}(\varepsilon_h)$ and $\mathrm{cov}(\varepsilon_\theta, \varepsilon_h)$ is estimated by equations (**??**) and (**??**) using the unbiased covariance for $\mathrm{cov}(\varepsilon_{h_i}, \varepsilon_{h_j})$.

Section **??** describes the optimal number of humans does not always exist. However, in reality there is an upper limit $N_{\max}$ depending on the number of respondents. Therefore, for the human-machine ensemble method, we used the n that minimizes the expected error in the range of $0 \leq n \leq N_{\max}$. In this case, unless $\mathrm{MSE}(Y_{\theta,h}(0)) = \mathrm{MSE}(Y_{\theta,h}(N_{\max}))$ holds, n is specified uniquely. When this equation holds, we set $n^* = N_{\max}$. The experiment fixed $N_{\max} = 5$, which is the same as the members sampled in the survey forecasts. Furthermore, when the expected error of humans only is less than that of combination of humans and a machine, the ensemble does not include the machine forecast.

1.2.5 Assessing forecasting methods

The test set starts from 2008. We also prepared the test set that starts from 1998 only for CPI. The reason is that the SPF has the short sample period since it started from 1981 and the Livingston Survey forecasts only CPI.

Table 1.2: The varieties of the ensembles.

	CPI-LIV	CPI-SPF	CoreCPI	PCE	CorePCE	CPI-6M	CPI-1998
target index	CPI	CPI	CoreCPI	PCE	CorePCE	CPI	CPI
survey source	LIV	SPF	SPF	SPF	SPF	LIV	LIV
forecast period	12M	12M	12M	12M	12M	6M	12M
dataset boundary	2008	2008	2008	2008	2008	2008	1998

The creation of RNN models and estimation of the ensemble parameters uses only the training set. The ARMA(1,1) model is recreated at each forecasting since the creation of the model is easier than RNN. That is, when forecasting annual inflation at 2008Q1, the creation of the model uses the samples up to 2007Q1, and when forecasting it at 2008Q2, the model is recreated using the samples up to 2007Q2. Thus, the creation of ARMA(1,1) models uses the data up to the time of forecasting.

The frequency of forecasts in the test set is quarterly as the same to the benchmark, ARMA(1,1). The number of samples in the test set that starts 2008 is 35 from 2008Q1 to 2016Q3. The Livingston Survey and the ensemble using it can only make forecasts on June and December, so the number of samples in the test set that starts 2008 is 17, and the one that starts 1998 is 37. Although the RNN models can make forecasts monthly, they forecast only on March, June, September and December.

Table 1.2 shows the all ensembles. There are seven different ensembles depending on the target index, the survey source, the period to forecast, and the boundary of the training and test set.

We assess forecast accuracy with the *Root Mean Squared Error (RMSE)*. When an actual inflation and forecast value at time t is y_i and $\hat{y}_i$ respectively, and forecasting M times, the RMSE is

$$\mathrm{RMSE}(\hat{\boldsymbol{y}}) = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(y_i - \hat{y}_i)}. \tag{1.10}$$

In addition, we also report the ratio of RMSEs relative to the benchmark, ARMA(1,1).