Transience prediction

July 29, 2021

1 Ideas

In time series literature, under a squared loss $L(\cdot)$, the conditional forecast, which turns out to be the frequently used standard multiple horizon forecast, is the best point forecast that minimizes the loss if the model is specified correctly. In other words, if the time series does experience a shock but the shock is transient, as $D_i \equiv T_i - T_i^* + 1 \to \infty$, the mean squared loss of the conditional forecast for D_i horizons should converge to the minimum of mean squared loss that any D_i-horizon forecast can achieve. On the other hand, if the time series do experience a shock but the shock is permanent and large, D_i-horizon conditional forecast adjusted by additive shock should enjoy a better loss than the unadjusted conditional forecast obtained from the training data that do not experience a shock. Essentially, the problem of judging whether a shock is permanent or transient boils down into comparing the losses of adjusted D_i-horizon conditional forecast and the unadjusted one.

Quaedvlieg [2021] proposes a test to jointly compare 1 to H-ahead forecasts for a sequence of models such that users can get a sense of the overall quality of the model. Note that D_i and H need not be equal. D_i is the horizon while H proxies the maximum requirements of the model predictive ability. In our case, this method can be applied to a paired comparison of adjusted D_i-horizon conditional forecast and unadjusted one. The methodology of Quaedvlieg [2021] involves a specification of a loss function $L(\cdot)$ that is used to compare the losses between two models. Moreover, the proposed test statistic is a function of those losses. In other words, it requires information of the responses that are to be forecasted. Therefore, a naive application of this method to compare adjusted D_i-horizon conditional forecast and the unadjusted one would fail. It is because from the perspective of practitioners, a retrospective comparison is not as useful as a prospective comparison that can predict whether the shock is transient or permanent without observing the shock and future series.

To deal with this problem, motivated by Abadie et al. [2010] and Lin and Eck [2020], it is possible to approximate the probability that the shock is transient using the donor pool under suitable assumptions. Then, we illustrate the idea as below. Let the donor pool size be n. For $i = 1, \ldots, n+1$ and $t = 1, \ldots, T_i$, we assume

$$
\begin{aligned}
y_{i,t} &= K(\mathcal{F}_{i,t-1}) + \mathbf{x}_{i,t}\boldsymbol{\beta} + \alpha_i I(t > T_i^*) + \varepsilon_{i,t}, \\
\alpha_i &= \mu_\alpha + \mathbf{x}_{i,T_i^*+1}\boldsymbol{\gamma}_i + \tilde{\varepsilon}_i
\end{aligned}
$$

where $\mathcal{F}_{i,t-1}$ denotes the information before $t-1$ of time series i, $K(\cdot)$ is a general real-valued function, $\mathbf{x}_{i,t}, \boldsymbol{\beta} \in \mathbb{R}^p$, $\boldsymbol{z}_{i,t}, \boldsymbol{\gamma}_i \in \mathbb{R}^q$, $\tilde{\varepsilon}_{i,t}$ and $\varepsilon_{i,t}$ follow some well-defined distribution with existing first and second moments, and T_i^* is the time point that the user knows a shock is about to come at $t = T_i^* + 1$. We further assume, $\tilde{\varepsilon}_{i,t}$ and $\varepsilon_{i,t}$ are independent. This implies that $\alpha_{i,t}$ and $\varepsilon_{i,t}$ are independent.

Note that $\|(\alpha_{i,t})_{t=T_i^*+1}^{t=T_i}\|_2$ should be close to zero if the shock is transient. Also, more general than Lin and Eck [2020], $\mathbf{z}_{i,t}$ is a vector of known covariates for the shock effects. Note that the model for $\alpha_{i,t}$ may need revision since specification of covariates for H time points in practice may be difficult, and involve large uncertainty and estimation error.

Let $\boldsymbol{d}_i^h = (d_{i,t}^h)_{t=1}^{T_i}$ be the loss differences of adjusted and unadjusted forecast for $i = 2, \ldots, n+1$ at the h-ahead forecast. Define $\mathbb{E}(d_{i,t}^h) = \mu_{i,t}^h$ and

$$\mu_i^h = \lim_{T_i \to \infty} \frac{1}{T_i} \sum_{t=1}^{T_i} \mu_{i,t}^h.$$

Note that in the case of post-shock prediction, before the shock time point, the adjusted forecast and unadjusted forecast are essentially the same. It implies

$$\mu_i^h = \lim_{T_i \to \infty} \frac{1}{T_i} \sum_{t=T_i^*+1}^{T_i} \mu_{i,t}^h.$$

In other words, in order not make the test trivial, we have to make sure $T_i = O(T_i - T_i^* + 1)$, i.e., $T_i - T_i^* + 1$is not negligible compared to T_i. Otherwise, μ_i^h will be always zero and there is no need for testing. Also,

Quaedvlieg [2021] proposes two tests for testing the following two hypotheses respectively,

$$\begin{aligned} H_{0,1}\colon \mu_i^{(\text{Avg})} \equiv \sum_{h=1}^{H} a_h \mu^h \leq 0 \quad \text{versus} \quad H_{1,1}\colon \mu_i^{(\text{Avg})} > 0 \\ H_{0,2}\colon \mu_i^{(\text{Unif})} \equiv \min_{h=1,\ldots,H} \mu_i^h > 0 \quad \text{versus} \quad H_{1,2}\colon \mu_i^{(\text{Unif})} \leq 0, \end{aligned}$$

where $\sum_{h=1}^{H} a_h = 1$ and a_h is typically selected to be $1/H$. The first hypothesis is for testing average predictive superiority whereas the second is for uniform superiority. From the definition, apparently, the second test is more stringent and more powerful but it is less likely to reject. It rejects only when the adjusted forecast is significantly better than the unadjusted one. Also, Quaedvlieg [2021] note that if H is large, the power of the test may decrease.

Consider a hypothesis test $\mathcal{T}_i$ that considers testing one of the two hypotheses. Quaedvlieg [2021] shows that his test is powerful as $T_i \to \infty$. So, we may define the permanence of the shock as

$$S_i = I(\mathcal{T}_i \text{ rejects } H_0) = I(\boldsymbol{\alpha}_i \text{ is permanent}).$$

We further assume that for $i = 1, \ldots, n+1$,

$$\mathbb{E}(S_i) = \mathbb{P}(\boldsymbol{\alpha}_i \text{ is permanent}) = s \text{ for some } s \in [0,1].$$

Suppose we can obtain a weighting $\mathbf{W} = (w_2, \ldots, w_{n+1})$. As a result, we can find an unbiased estimate for $\mathbb{P}(\boldsymbol{\alpha}_1 \text{ is permanent})$ with

$$\hat{\mathbb{P}}(\boldsymbol{\alpha}_1 \text{ is permanent}) = \sum_{i=2}^{n+1} w_i \cdot S_i$$
$$\mathbb{E}(\hat{\mathbb{P}}(\boldsymbol{\alpha}_1 \text{ is permanent})) = \sum_{i=2}^{n+1} w_i \cdot \mathbb{E}(S_i) = \sum_{i=2}^{n+1} w_i \cdot s = s.$$

We may construct the weighting using synthetic control method as below. For $i = 1, \ldots, n+2$, define

$$\mathbf{Z}_i = \begin{pmatrix} \mathbf{z}_{i,T_i^*+1} \\ \vdots \\ \mathbf{z}_{i,T_i^*+H} \end{pmatrix} \quad \text{and} \quad \mathcal{W} = \{\mathbf{W} \in [0,1]^n : \mathbf{1}_n' \mathbf{W} = 1\}$$

Suppose there exists $\mathbf{W}^* \in \mathcal{W}$ with $\mathbf{W}^* = (w_2^*, \ldots, w_{n+1}^*)$ such that

$$\mathbf{Z}_1 = \sum_{i=2}^{n+1} w_i^* \mathbf{Z}_i.$$

We may estimate $\mathbf{W}^*$ as

$$\mathbf{W}^* = \underset{\mathbf{W} \in \mathcal{W}}{\arg\min} \, \|\mathbf{Z}_1 - \sum_{i=2}^{n+1} w_i^* \mathbf{Z}_i\|_2.$$

2 Definitions

2.1 stochastic process and dynamic system

Definition 1 (stochastic process)**.** A stochastic process $\{Y_t\}_{t \in T}$ is a collection of random variables Y_t, taking values in a common measurable space $(\Xi, \mathcal{X})$, indexed by a set T

Definition 2 (one-parameter process)**.** A process whose index set T has one dimension is a one-parameter process.

Definition 3 (dynamical system)**.** A dynamical system consists of a measurable space Ξ, a σ-algebra $\mathcal{X}$ on Ξ, a probability measure μ defined on $\mathcal{X}$, and a mapping $T : \Xi \mapsto \Xi$ which is $(\mathcal{X}, \mathcal{X})$-measurable

2.2 mixing

Definition 4 (mixing)**.** A dynamical system $(\Xi, \mathcal{X}, \mu, T)$ is mixing when, for any $A, B \in \mathcal{X}$,

$$\lim_{t\to\infty} |\mu(A \cap T^{-t}(B)) - \mu(A)\mu(T^{-t}(B))| = 0,$$

where $T^{-t}(A) = T^{-1}(T^{-t+1}(A))$. For stochastic process, "mixing" means "asymptotically independent", which means the statistical dependence between $Y(t_1)$ and $Y(t_2)$ goes to zero as $|t_1 - t_2|$ increases.

Definition 5 (near epoch dependence)**.** $\{Y_t\}$ is a near epoch dependent (NED) on a mixing process $\{V_t\}$ if $\mathbb{E}(Y_t^2) < \infty$ and $v_k = \sup_t \left\| Y_t - \mathbb{E}_{t-k}^{t+k}(Y_t) \right\|_2 \to 0$ as $k \to \infty$, where $\|\cdot\|_p$ is the L_p norm and $\mathbb{E}_{t-k}^{t+k}(\cdot) \equiv \mathbb{E}(\cdot|\mathcal{F}_{t-k}^{t+k})$, where $\mathcal{F}_{t-k}^{t+k} \equiv \sigma(V_{t-k}, \ldots, V_{t+k})$ is the σ-algebra generated by $V_{t-k}, \ldots, V_{t+k}$.

Definition 6. If $v_k = O(k^{-a-\delta})$ in Definition 5 for some $\delta > 0$, we say $\{Y_t\}$ is NED of size $-a$.

Definition 7 (conditional stationarity)**.** A one-parameter process $\{Y_t\}$ is conditionally stationary if for all $n \in \mathbb{N}$ and every set of $n+1$ indices $t_1, \ldots, t_{n+1} \in T$, $t_i < t_{i+1}$, and every shift τ,

$$\mathcal{L}(Y_{t_{n+1}}|Y_{t_1}, Y_{t_2}, \ldots, Y_{t_n}) = \mathcal{L}(Y_{t_{n+1}+\tau}|Y_{t_1+\tau}, Y_{t_2+\tau}, \ldots, Y_{t_n+\tau}) \quad a.s.,$$

where $\mathcal{L}(Y|X)$ is the distribution function of a Y conditional on X.

Definition 8 (α-mixing)**.** For a stochastic process Y_t, the α-mixing coefficient is

$$\alpha(t_1, t_2) = \sup\{|\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)| \colon A \in \sigma(Y_{t_1}^-), B \in \sigma(Y_{t_2}^+)\},$$

where $Y^+ = \max\{Y, 0\}$ and $Y^- = \max\{-Y, 0\}$. If the system is conditionally stationary, $\alpha(t_1, t_2) = \alpha(t_2, t_1) = \alpha(|t_1 - t_2|) \equiv \alpha(\tau)$. If $\alpha(\tau) \to 0$ as $\tau \to \infty$, the process is α-mixing.

3 Simulation Setup

Let n denote the donor pool size, p denote the number of covariates used, H denote the number of horizon used, T_i denote the length of time series to be evaluated for time series i, K_i denote the training sample size used for each forecasting time series i, T_i^* denote the time point just before the realization of the shock for time series i for $i = 1, \ldots, n+1$.

n, p, and H are pre-determined. $T_i, K_i \sim \text{Gamma}(15, 10)$. The total sample size for ith time series is $T_i + K_i + H$. T_i^* is randomly sampled from $\lceil \frac{1}{4} T_i \rceil + 1$ to $\lceil \frac{3}{4} T_i \rceil + K_i + H$. If $T_i, K_i < 90$, we force them to be 90. The adopted model for the data is as below:

$$
\begin{aligned}
y_{i,t} &= \eta_i + \phi_i y_{i,t-1} + \mathbf{x}_{i,t} \boldsymbol{\beta}_i + \alpha_i I(t > T_i^*) + \varepsilon_{i,t}, \\
\alpha_i &= \mu_\alpha + \mathbf{x}_{i,T_i^*+1} \boldsymbol{\gamma}_i + \tilde{\varepsilon}_i,
\end{aligned}
$$

where

$$
\begin{aligned}
\phi_i &\sim \text{ indep. } U(0,1) \\
\eta_i &\sim \text{ indep. } \mathcal{N}(0,1) \\
\varepsilon_{i,t} &\sim \text{ indep. } \mathcal{N}(0, \sigma^2) \\
\tilde{\varepsilon}_i &\sim \text{ indep. } \mathcal{N}(0, \sigma_\alpha^2) \\
\boldsymbol{\gamma}_i &\sim \text{ indep. } \mathcal{N}(\mu_\gamma \mathbf{1}_p, \sigma_\gamma^2 \mathbf{I}_p) \\
\boldsymbol{\beta}_i &\sim \text{ indep. } \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p).
\end{aligned}
$$

Moreover, the elements of $\mathbf{x}_{i,t}$ are independently distributed as $\text{Gamma}(1, \delta)$.

Note that K_i is training sample size for time series i. Consider

$$
\begin{aligned}
K_i &\sim \lceil \text{Gamma}(a_K, b_K) \rceil \\
T_i &\sim \lceil \text{Gamma}(a_T, b_T) \rceil \\
T_i^* &\equiv \lceil 0.5 \cdot T_i \rceil,
\end{aligned}
$$

Then, we consider the following simulation setup

```
ns <- c(5, 10, 20, 40)
Tscale <- Kscale <- 1 / 2 # b_T, b_K
K.T.shape <- c(200, 400, 800, 1600) # for K_i and T_i
mu.gamma.delta <- 2 # mean for parameter vector of shock
sigma.delta.gamma <- 0.1 # sd for parameter vector of shock
sigma.alpha <- 0.05 # sd for shock noise
sigma <- 0.1 # sd for response noise
mu.alpha <- 50 # intercept for shock (relatively large)
H <- 8
ell <- 4
scale <- 2 # scale for covariates that follow Gamma distribution
```

$$
\begin{aligned}
y_{i,t} &= \eta_i + \phi_i y_{i,t-1} + \mathbf{x}_{i,t} \boldsymbol{\beta}_i + \xi_i \cdot I(t > T_i^*) + \varepsilon_{i,t}, \\
\xi_i &= \alpha_i \cdot e^{-(t - T_i^* - 1)} \\
\alpha_i &= \mu_\alpha + \mathbf{x}_{i,T_i^*+1} \boldsymbol{\gamma}_i + \tilde{\varepsilon}_i,
\end{aligned}
$$

Table 1: 50 MC simulations with varying n and σ_α ($B = 200$, $H = 12$, $\ell = 3$, $\mu_\alpha = 10$) MC mean and standard errors for absolute differences between p_1 and estimated p_1.

n	σ_α	$\lvert\hat{p} - p_1\rvert$	$\lvert\hat{I} - I_1\rvert$
5	1	0.083 (0.028)	0.084 (0.028)
	5	0.083 (0.028)	0.084 (0.028)
	10	0.083 (0.028)	0.084 (0.028)
	25	0.074 (0.027)	0.074 (0.027)
	100	0.053 (0.019)	0.053 (0.019)
10	1	0.063 (0.03)	0.063 (0.03)
	5	0.063 (0.03)	0.063 (0.03)
	10	0.063 (0.03)	0.063 (0.03)
	25	0.063 (0.03)	0.063 (0.03)
	100	0.052 (0.025)	0.052 (0.025)
15	1	0.065 (0.026)	0.066 (0.026)
	5	0.065 (0.026)	0.066 (0.026)
	10	0.065 (0.026)	0.066 (0.026)
	25	0.065 (0.026)	0.066 (0.026)
	100	0.087 (0.029)	0.087 (0.029)
25	1	0.044 (0.021)	0.044 (0.021)
	5	0.043 (0.021)	0.044 (0.021)
	10	0.05 (0.022)	0.05 (0.022)
	25	0.05 (0.022)	0.05 (0.022)
	100	0.048 (0.022)	0.049 (0.022)

Table 2: 100 MC simulations with varying σ and σ_α ($B = 200$, $H = 12$, $\ell = 3$, $n = 10$) MC mean and standard errors for absolute differences between p_1 and estimated p_1.

σ	σ_α	$\lvert\hat{p} - p_1\rvert$	$\lvert\hat{I} - I_1\rvert$
5	5	0.039 (0.016)	0.04 (0.016)
	10	0.039 (0.016)	0.04 (0.016)
	25	0.054 (0.019)	0.054 (0.019)
	50	0.059 (0.019)	0.059 (0.019)
	100	0.098 (0.026)	0.098 (0.026)
10	5	0.044 (0.016)	0.044 (0.016)
	10	0.044 (0.016)	0.044 (0.016)
	25	0.042 (0.016)	0.042 (0.016)
	50	0.065 (0.019)	0.065 (0.019)
	100	0.101 (0.026)	0.101 (0.026)
25	5	0.045 (0.018)	0.046 (0.018)
	10	0.058 (0.02)	0.059 (0.02)
	25	0.06 (0.02)	0.061 (0.02)
	50	0.105 (0.026)	0.106 (0.026)
	100	0.122 (0.027)	0.122 (0.027)
50	5	0.075 (0.023)	0.076 (0.023)
	10	0.085 (0.025)	0.086 (0.025)
	25	0.08 (0.023)	0.08 (0.023)
	50	0.14 (0.029)	0.14 (0.029)
	100	0.148 (0.027)	0.148 (0.027)
100	5	0.148 (0.029)	0.149 (0.029)
	10	0.151 (0.029)	0.151 (0.029)
	25	0.162 (0.029)	0.162 (0.029)
	50	0.229 (0.034)	0.229 (0.034)
	100	0.272 (0.034)	0.272 (0.034)

Table 3: 100 MC simulations for decaying shock effects with varying μ_α and H ($B = 200$, $\mu_\gamma = 2$, $\ell = 3$, $n = 10$)

μ_α	H	$\lvert\hat{p} - p_1\rvert$	$\lvert\hat{I} - I_1\rvert$	Mean of p-values in the donor pool
0	2	0.11 (0.023)	0.114 (0.023)	0.08 (0.009)
5	2	0.1 (0.021)	0.1 (0.021)	0.079 (0.009)
50	2	0.075 (0.018)	0.075 (0.018)	0.062 (0.008)
100	2	0.064 (0.017)	0.064 (0.017)	0.048 (0.007)
0	4	0.12 (0.025)	0.124 (0.025)	0.069 (0.007)
5	4	0.12 (0.025)	0.12 (0.025)	0.064 (0.007)
50	4	0.087 (0.022)	0.094 (0.023)	0.046 (0.007)
100	4	0.049 (0.016)	0.049 (0.017)	0.033 (0.006)
0	8	0.184 (0.031)	0.192 (0.032)	0.07 (0.008)
5	8	0.181 (0.032)	0.191 (0.033)	0.067 (0.008)
50	8	0.139 (0.029)	0.139 (0.029)	0.05 (0.007)
100	8	0.102 (0.024)	0.103 (0.024)	0.04 (0.007)
0	16	0.202 (0.031)	0.202 (0.031)	0.097 (0.008)
5	16	0.202 (0.031)	0.202 (0.031)	0.095 (0.008)
50	16	0.168 (0.031)	0.168 (0.031)	0.064 (0.008)
100	16	0.125 (0.029)	0.125 (0.029)	0.044 (0.006)

Table 4: 100 MC simulations for decaying shock effects with varying ℓ and H ($B = 200$, $\mu_\gamma = 2$, $\mu_\alpha = 5$, $n = 10$)

ℓ	H	$\lvert\hat{p} - p_1\rvert$	$\lvert\hat{I} - I_1\rvert$	Mean of p-values in the donor pool
2	2	0.107 (0.026)	0.12 (0.028)	0.064 (0.008)
4		0.108 (0.022)	0.118 (0.023)	0.077 (0.008)
8		0.105 (0.024)	0.107 (0.024)	0.067 (0.009)
16		0.12 (0.025)	0.124 (0.026)	0.07 (0.009)
2	4	0.162 (0.032)	0.172 (0.033)	0.061 (0.008)
4		0.089 (0.017)	0.097 (0.019)	0.08 (0.008)
8		0.135 (0.027)	0.135 (0.027)	0.08 (0.009)
16		0.156 (0.031)	0.158 (0.031)	0.063 (0.007)
2	8	0.108 (0.025)	0.112 (0.025)	0.079 (0.009)
4		0.134 (0.029)	0.136 (0.029)	0.071 (0.008)
8		0.186 (0.033)	0.186 (0.032)	0.075 (0.008)
16		0.125 (0.026)	0.125 (0.026)	0.071 (0.008)
2	16	0.222 (0.033)	0.225 (0.033)	0.094 (0.01)
4		0.124 (0.024)	0.126 (0.024)	0.081 (0.008)
8		0.14 (0.028)	0.14 (0.028)	0.075 (0.008)
16		0.162 (0.032)	0.162 (0.032)	0.068 (0.009)

4 Conditions

1. Donor pool sample size $n \to \infty$
2. $\|\alpha_1 - \hat{\alpha}_w\|_2 \to 0$ as $n \to \infty$ (this can be realized by setting σ_α small enough and $\text{Var}(\gamma)$ small enough)
3. The set of weights that are positive is fixed and finite asymptotically
4. Conditions of Quaedvlieg and $B \to \infty$
5. $T_i \to \infty$ for $i = 2, \ldots, n+1$ and $T_i^* = o(T_i)$.

If all those conditions hold, $\|p_1 - \hat{p}\| \to 0$.

References

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American Statistical Association*, 105(490):493–5ARTICLE05, 2010.

Jilei Lin and Daniel J Eck. Minimizing post-shock forecasting error through aggregation of outside information. *arXiv preprint arXiv:2008.11756*, 2020.

Rogier Quaedvlieg. Multi-horizon forecast comparison. *Journal of Business & Economic Statistics*, 39(1):40–53, 2021.