Time-varying Forecast Combination for High-Dimensional Data

Bin Chen and Kenwin Maung

University of Rochester

October, 2021

The contributions of our paper

- When the number of forecasts is finite, we consider a new estimator for time-varying forecast combination weights and study its asymptotic properties. The estimator can be viewed as a generalization of the classical Granger and Ramanathan's (1984) regression estimator.
- We study a data-driven bandwidth selection method. It can be viewed as an analogue of the window size selection for classical rolling regression.
- When the number of potential forecasts is allowed to be larger than the sample size, we consider a two-step procedure to select the significant forecasts and estimate time-varying combination weights.
- The penalized estimator is shown to have the oracle property.

Time-varying Combination Weights

- Assume that a decision maker is interested in predicting some univariate series y_{t+1}, conditional on I_t, the information at time t, which consists a set of individual forecasts $f_t = (f_{t1,} f_{t2}, ..., f_{td})^{\mathsf{T}}$ in addition to current and past values of y, i.e., $I_t = (y_s, f_s)_{s=1}^{t}$
- We consider forecast combination with time-varying weights:

$$y_{t+1} = \omega_{0t} + f_t^{\mathsf{T}} \omega_{1t} + \varepsilon_{t+1}, \qquad t = 1, ..., T,$$

where $(\omega_{0t}, \omega_{1t}^{\mathsf{T}})^{\mathsf{T}}$ are adapted to the current information set I_t.

Nonparametric estimation

- We adopt the framework of the smooth time-varying regression model proposed by Robinson(1989) and Cai (2007).
- Assume $\omega_{0t} = \omega_0\left(t/T\right)$ and $\omega_{1t} = \omega_1\left(t/T\right)$ are smooth function of the standardized time t/T over [0,1].:

$$\begin{aligned} y_{t+1} &= \omega_0\left(t/T\right) + f_t^{\intercal}\omega_1\left(t/T\right) + \varepsilon_{t+1} \\ &\equiv x_t^{\intercal}\beta\left(t/T\right) + \varepsilon_{t+1} \\ &\simeq x_t^{\intercal}\beta_s^0 + \frac{s-t}{T}x_t^{\intercal}\beta_s^1 + \varepsilon_{t+1}, \end{aligned}$$

where s is in some neighborhood of t.

Nonparametric estimation

Consider the following local least squares problem:

$$\min_{\gamma \in \mathbb{R}^{2(d+1)}} \sum_{s=t-\lfloor Th \rfloor, s \neq t}^{t+\lfloor Th \rfloor} k\left(\frac{s-t}{Th}\right)\left[y_{s+1} - \alpha_0^\intercal x_s - \alpha_1^\intercal \left(\frac{s-t}{T}\right) x_s\right]^2$$

$$= \sum_{s=t-\lfloor Th \rfloor, s \neq t}^{t+\lfloor Th \rfloor} k_{st}\left(y_{s+1} - \gamma^\intercal q_{st}\right)^2,$$

where $\gamma = (\alpha_0^\intercal, \alpha_1^\intercal)^\intercal$ is a $2(d+1) \times 1$ vector, α_j is a $(d+1) \times 1$ coefficient vector for $(\frac{s-t}{T})^j x_s$, $j = 0, 1$, $q_{st} = z_{st} \otimes x_s$ is a $2(d+1) \times 1$ vector, and $\otimes$ is the Kronecker product.

Window Length Selection

- Global Leave-one-out cross-validation (CV)

$$\hat{h}_{CV} = \arg \min_{c_1 T^{-1/5} \leq h \leq c_2 T^{-1/5}} CV(h),$$

  where

$$CV(h) = \Sigma_{s=1}^{T} (y_{s+1} - x_s^{\intercal} \hat{\beta}_s)^2.$$

- Theorem 1: Suppose Assumptions A.1-A.4 hold. Then

$$CV(h) = IMSCFE(h) + o_P\left(T^{-4/5}\right)$$

  uniformly in h.
- Proposition 2: Suppose Assumptions A.1-A.4 hold. As $T \to \infty$,

$$\hat{h}/h^{opt} \to 1, \ a.s.$$

where h^{opt} is the optimal bandwidth which minimizes $IMSCFE(h)$.

High Dimensional Case

- In practice, there might be very large number of forecasts and some forecasts might be redundant. So we consider the case with ultra-high dimensional forecasts

$$y_{t+1} = \omega_{0t} + f_t^{\intercal}\omega_{1t} + \varepsilon_{t+1}, \qquad t = 1, ..., T,$$

  where the f_t is $p_T \times 1$ and p_T can be larger than T.
- We assume that $1 \leq d_T < p_T$ such that $\omega_{1t,j} \neq 0$ for $1 \leq j \leq d_T$ and $\omega_{1t,j} = 0$ for $d_T < j \leq p_T$.
-

  d_T may diverge with the sample size T but it is much smaller than T ar dimension of the whole forecasts p_T.

Penalized Local Linear Estimation

- With high dimension, the local linear estimation doesn't work. We first consider penalized local linear estimation

$$\tilde{\gamma}_t = \arg\min_{\gamma} \sum_{s=t-\lfloor Th \rfloor, s \neq t}^{t+\lfloor Th \rfloor} k\left(\frac{s-t}{Th}\right)\left[y_{s+1} - \alpha_0^{\intercal} x_s \right.$$
$$\left. -\alpha_1^{\intercal}\left(\frac{s-t}{T}\right) x_s\right]^2 + \lambda_1 \left|\alpha_0\right| + \lambda_2 \left|h\alpha_1\right|,$$

where λ_1 and λ_2 are two tuning parameters.

- Then the local linear estimator for β_t is given by

$$\tilde{\beta}_t = (e_1^{\intercal} \otimes I_{(p_T+1)})\tilde{\gamma}_t.$$

Penalized Estimation with Group SCAD

- Further consider the estimation with group scad

$$\hat{\Upsilon}^h = \underset{\Upsilon \in \mathbb{R}^{T\times(p_T+1)}}{\arg\min} T^{-1} \sum_{t=1}^{T-1} \sum_{s=t-\lfloor Th \rfloor, s\neq t}^{t+\lfloor Th \rfloor} k_{st} [y_{s+1} - \alpha_{0t}^{\mathsf{T}} x_s - \alpha_{1t}^{\mathsf{T}} \times \left(\frac{s-t}{T}\right) x_s]^2 + \sum_{j=1}^{p_T} p_{\lambda_3}^{'} \left(\|\tilde{B}_j\|\right) \|\alpha_{0,j}\| + \sum_{j=1}^{p_T} p_{\lambda_4}^{\prime} \left(\tilde{D}_j\right) \|h\alpha_{1,j}\|$$

where λ_3 and λ_4 are two tuning parameters, $\alpha_i = (\alpha_{i1}, \alpha_{i2}, ..., \alpha_{iT})^{\mathsf{T}}$ for $i = 0, 1$, and $\alpha_{i,j}$ is the *jth* column of α_i, $\tilde{B}_j$ is the *jth* column of the LASSO-based local linear estimator $\tilde{B}$.

Penalized Estimation with Group SCAD

- $$\tilde{D}_j = \left\{ \sum_{t=1}^{T-1} \left[\tilde{\beta}_{t,j} - \frac{1}{T} \sum_{s=1}^{T} \tilde{\beta}_{s,j} \right]^2 \right\}^{1/2},$$

  which measures the smoothness of the LASSO-based local linear estimator $\tilde{B}$. Moreover, $p'_\lambda(\cdot)$ is the derivative of the SCAD penalty function with regularization parameter λ defined by

  $$p'_\lambda(x) = \lambda \left[1(x \leq \lambda) + \frac{(a\lambda - x)_+}{(a-1)\lambda} 1(x > \lambda) \right].$$

- Then the local linear estimator for β_t with the group SCAD penalty is given by

  $$\hat{\beta}_t^h = (e_1^\intercal \otimes I_{(p_T+1)}) \hat{\gamma}_t^h,$$

Penalized Estimation with Group SCAD

- Theorem 2**:** *Suppose Assumptions A.1,A.2,A.3(i), A.4, HD.1–3 hold. We have*

$$P(\hat{S} = S_0) \to 1,$$

as $T \to \infty$.

- Theorem 3: *Suppose Assumptions A.1,A.2,A.3(i), HD.1–3 hold. We have*

$$\sqrt{Th} A_T \Omega^{o-1/2}(\tau) \left\{ \hat{\beta}^{o,h}(\tau) - \beta^{o}(\tau) - \frac{h^2}{2}\mu_2 \beta^{o\prime\prime}(\tau) + o_P\left(h^2\right) \right.$$
$$\to {}^{d} N(0, G),$$

as $T \to \infty$, where $\Omega^{o}(\tau) = 2\nu_0 M^{o-1}(\tau) V^{o}(\tau) M^{o-1}(\tau)$, and A_T is an arbitrary $q \times d$ matrix such that $A_T A^{\intercal} \to G$ for a given finite q.