Machine learning time series regressions with an application to nowcasting

Andrii Babii[1] Eric Ghysels[1] Jonas Striaukas[2]

[1]UNC Chapel Hill

[2]UC Louvain

ASSA 2021

Based on:

- Andrii Babii, Eric Ghysels, and Jonas Striaukas (2020) **Machine Learning Time Series Regressions with an Application to Nowcasting**, arXiv:2005.14057
- Andrii Babii, Eric Ghysels, and Jonas Striaukas (2020) **Inference for High-dimensional Regressions with Heteroskedasticity and Autocorrelation**, arXiv:1912.06307
- Andrii Babii, Eric Ghysels, and Jonas Striaukas (2020) **Machine learning panel data regressions with an application to nowcasting price earnings ratios**, arXiv:2008.03600

Motivation

1. Nowcasting/forecasting in data-rich environment with high-dimensional real-time data
2. HAC-based post model selection inference: Granger causality tests
3. Examples: nowcasting quarterly US GDP growth/corporate earnings/volatilities with macroeconomic, financial, and textual news data measured potentially at higher frequencies

Related literature

LASSO

Chernozhukov, Härdle, Huang, and Wang (2020), Belloni, Chernozhukov, and Hansen (2014), van de Geer, Bühlmann, Ritov, and Dezeure (2014), Simon, Friedman, Hastie, and Tibshirani (2013), Bickel, Ritov, and Tsybakov (2009), Tibshirani (1996).

MIDAS projections

Andreou, Ghysels, and Kourtellos (2013), Ghysels, Santa-Clara, and Valkanov (2006), Mogliani and Simoni (2020).

Dynamic factor models

Forni, Hallin, Lippi, and Reichlin (2000), Stock and Watson (2002), Gianone, Reichlin and Small (2008), Banbura, Giannone, Modugno, and Reichlin (2013).

Contributions

New methodology

Nowcasting/forecasting in data-rich environments: sparse group LASSO (sg-LASSO) + MIDAS with Legendre polynomials.

New concentration inequalities

Fuk-Nagaev inequalities for heavy-tailed τ-mixing processes.

Theoretical guarantees for heavy-tailed data

Oracle inequalities, debiased CLT with explicit bias correction, and convergence rate of the HAC estimator based on sg-LASSO residuals.

Application to nowcasting US GDP growth:

- either superior or comparable to NY FED nowcasts at different horizons on the same dataset of macroeconomic data;
- additional gains with textual news and financial data.

Table of Contents

High-dimensional (mixed-frequency) time series

1. K predictors, each measured at $j \in [m] \triangleq \{1, 2, \ldots, m\}$ (high-frequency if $m > 1$) over T time periods

$$\{x_{t-(j-1)/m,k},\ k \in [K], j \in [m], t \in [T]\}.$$

2. Predictive regressions/projections

$$y_{t+h} = \mu + \sum_{k=1}^{K} \frac{1}{m} \sum_{j=1}^{m} \beta_k^{(j)} x_{t-(j-1)/m,k} + u_t.$$

3. High-dimensional problem: $m \times K$ slope parameters, where K and m can be large.

High-dimensional (mixed-frequency) time series

Dimensionality reduction

1. Parametrize slope coefficients via MIDAS approach

$$\beta_k^{(j)} = \omega\left(\frac{j-1}{m}; \beta_k\right), \qquad j \in [m]$$

   for some weight function $\omega : [0,1] \times \mathbf{R}^L \to \mathbf{R}$.

2. Sparse approximation to MIDAS weight function

$$\omega(t; \beta_k) \approx \sum_{l=1}^{L} \beta_{k,l} w_l(t),$$

   where $(w_l)_{l=1}^L$ is some dictionary and $\beta_k = (\beta_{k,1}, \dots, \beta_{k,L})$ is approximately sparse.

High-dimensional (mixed-frequency) time series

Choice of dictionary

1. orthogonal Legendre polynomials is one possibility: universal approximation for any continuous function.
2. Preferred to non-orthogonalized Almon (algebraic) polynomial due to better numerical properties: reduced multicolinearity and stable coefficients.
3. Includes constant $\implies$ averaging lags is a special case.
4. Linear in parameters regression: convex optimization.

Sparse-group LASSO regression/projection

Estimator

1. sg-LASSO solves

$$\operatorname{argmin}_{b\in\mathbf{R}^p}\|\mathbf{y}-\mathbf{X}b\|_T^2+2\lambda\Omega_\gamma(b)$$

with regularization parameter $\lambda \geq 0$ and penalty

$$\Omega_\gamma(b)=\gamma|b|_1+(1-\gamma)\|b\|_{2,1},$$

where $\gamma\in[0,1]$ is a relative weight, $|.|_1$ is ℓ_1 norm, $\|.\|_{2,1}$ is the group-LASSO norm.

2. Easy and fast to compute via proximal gradient descent.

Structured sparsity

Groups

1. For a single covariate a group is defined as a set of all its high-frequency lags;
2. High-dimensional model selection:
   - size of the dictionary?
   - selection of covariates?
3. Sparse-group LASSO performs selection at two levels
   - within groups: learning the shape of the MIDAS weight;
   - between groups: learning the most relevant covariates.
4. Covers LASSO and group LASSO as special cases.

sg-LASSO and dependent data

Weak dependence coefficients

1. τ-mixing coefficient of a stationary process $\xi_t \in \mathbf{R}$ with history $\mathcal{F}_t$

$$\tau_k = \sup_{j \geq 1} \frac{1}{j} \sup_{t+k \leq t_1 < \cdots < t_j} \tau(\mathcal{F}_t, (\xi_{t_1}, \dots, \xi_{t_j})),$$

   where $\tau(\mathcal{F}, \zeta) = \mathbb{E}\left|\sup_{f \in \mathrm{Lip}_1} |\mathbb{E}(f(\zeta)|\mathcal{F}) - \mathbb{E}(f(\zeta))|\right|$.

2. Dedecker-Prieur coupling: there exists measurable $\zeta^* =_d \zeta$ with $\zeta^* \perp\!\!\!\perp \mathcal{F}$ such that

$$\|\zeta^* - \zeta\|_1 = \tau(\mathcal{F}, \zeta).$$

3. Comparison to mixingale and α-mixing coefficient

$$\gamma_k \leq \tau_k \leq 2 \int_0^{2\alpha_k} Q(u) \mathrm{d}u,$$

   where γ_k is a mixingale coefficient and Q is the generalized inverse of $x \mapsto \Pr(|\xi_0| > u)$.

sg-LASSO and dependent data

New Fuk-Nagaev type inequality

Theorem

Suppose that $(\xi_t)_{t\in\mathbf{Z}}$ *is a centered stochastic process in* $\mathbf{R}^p$ *such that each coordinate* $j \in [p]$ *has*

(i) *tails:* $\sup_{t,j} \mathbb{E}|\xi_{t,j}|^q = O(1)$ *with* $q > 2$*;*

(ii) τ*-mixing coefficients:* $\tau_k \leq dk^{-a}$ *with* $a > (q-1)/(q-2)$*.*

Then for every $u > 0$

$$\Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T}\xi_t\right|_\infty > u\right) \leq \frac{c_1 p}{u^\kappa T^{\kappa-1}} + 4pe^{-c_2 Tu^2}.$$

sg-LASSO and dependent data

New Fuk-Nagaev inequality: comments

$$\Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T}\xi_t\right|_\infty > u\right) \leq \frac{c_1 p}{u^\kappa T^{\kappa-1}} + 4pe^{-c_2 Tu^2}$$

1. polynomial and exponential bound on tails of high-dimensional means, cf. Fuk and Nagaev (1971) for independent data;
2. sharp unlike Markov's + Rosenthal's bounds;
3. Dependence-tail exponent $\kappa = \frac{(a+1)q-1}{a+q-1} \xrightarrow[a\to\infty]{} q$.

sg-LASSO and dependent data

Convergence rates

Theorem

Under tail/dependence conditions and bounded restricted eigenvalue of $\Sigma = \mathbb{E}[\mathbf{X}^\top \mathbf{X}/T]$

$$|\hat{\beta} - \beta|_1 = O_P\left(\frac{s_\gamma p^{2/\kappa}}{T^{2-2/\kappa}} \vee s_\gamma \sqrt{\frac{\log p}{T}}\right)$$

$$\|\mathbf{X}(\hat{\beta} - \beta)\|_T^2 = O_P\left(\frac{s_\gamma p^{1/\kappa}}{T^{2-2/\kappa}} \vee s_\gamma \frac{\log p}{T}\right).$$

Comments:

1. OLS/QMLE requries $p/T = o(1)$, while we need $p/T^{\kappa-1} = o(1)$;
2. $\kappa > 2$ reflects the effect of tails, and dependence: lighter tails and less persistence $\implies$ can handle larger p.
3. Sparsity helps through s_γ, but is not needed.

HAC-based inference

Debiased CLT

Under some additional conditions for a group of coefficients $G \subset [p]$

$$\sqrt{T}(\hat{\beta}_G + b_G - \beta_G) \xrightarrow{d} N(0, \Xi_G).$$

- bias correction $b_G = \hat{\Theta}_G \mathbf{X}^\top(\mathbf{y} - \mathbf{X}\hat{\beta})/T$, where $\hat{\Theta}$ estimates $\Theta = \Sigma^{-1}$.
- long run variance: $\Xi_G = \lim_{T\to\infty} \mathrm{Var}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} u_t \Theta_G X_t\right)$.
- optimal bandwidth for the HAC estimator of Ξ_G depends on p.

Empirical illustration: nowcasting US GDP growth

Questions

1. How do high-dimensional ML models compare to simple benchmarks for nowcasting GDP growth in real time?
2. How do they compare to central bank nowcasts, e.g. New York Fed - using state space models.
3. Are non-standard textual news data useful for nowcasting?

Nowcasting US GDP growth

Data and setup

1. 34 macro series used in the NY FED nowcasting model (comparison of methods) and
   - 76 textual news series; see Bybee, Kelly, Manela, and Xiu (2020);
   - 8 financial series; see Andreou, Ghysels, and Kourtellos (2013).
2. Sample covering 1988Q1-2017Q2.
3. quarterly nowcasting for 2002Q1-2017Q2.
4. 5-fold cross-validation to select (λ, γ), where folds are adjacent over time blocks.

sg-LASSO vs. NY Fed (same data)

	Rel-RMSE	DM-stat-1	DM-stat-2
	2-month horizon		
AR(1)	2.056	0.612	2.985
sg-LASSO	0.739	-2.481	
NY Fed	0.946		2.481
	1-month horizon		
AR(1)	2.056	2.025	2.556
sg-LASSO	0.725	-0.818	
NY Fed	0.805		0.818
	End-of-quarter		
AR(1)	2.056	2.992	3.000
sg-LASSO	0.701	-0.077	
NY Fed	0.708		0.077
	p-value of aSPA test		
			0.046

Table: Nowcasting with macro data. RMSE relative to AR(1) benchmark. DM = Diebold Mariano statistics. aSPA = Average Superior Predictive Ability test over three horizons

Additional textual news and financial data

	Rel-RMSE	DM-stat-1	DM-stat-2
	2-month horizon		
Elastic Net	0.907	-0.266	0.976
sg-LASSO	0.779	-2.038	
NY Fed	0.946		2.038
	1-month horizon		
Elastic Net	0.990	1.341	2.508
sg-LASSO	0.672	-1.426	
NY Fed	0.805		1.426
	End-of-quarter		
Elastic Net	0.947	2.045	2.034
sg-LASSO	0.696	-0.156	
NY Fed	0.707		0.156
	p-value of aSPA test		
			0.042

Table: Nowcast with additional textual news and financial data. RMSE relative to AR(1) model. DM = Diebold Mariano statistics.

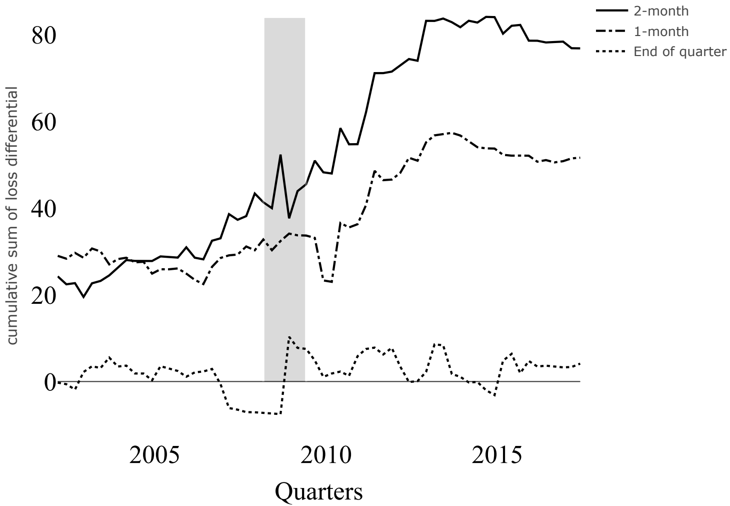

Figure: Cumulative sum of loss differentials of sg-LASSO-MIDAS model vs. New York Fed for three horizons.

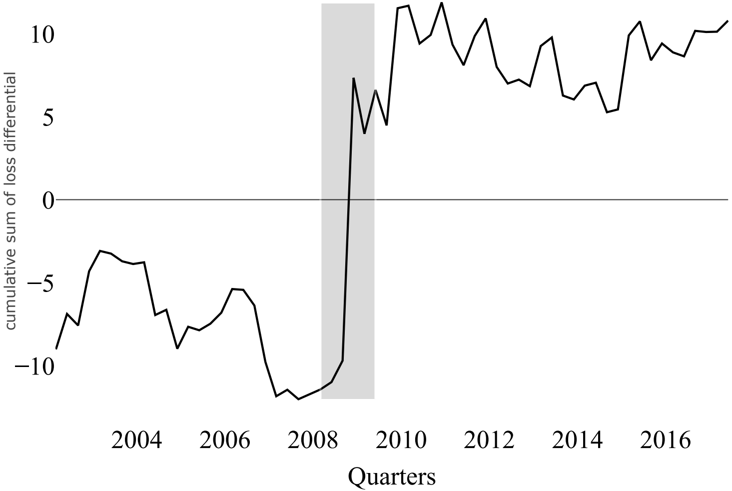

Figure: Cumulative sum of loss differentials of sg-LASSO-MIDAS nowcasts when we include vs. when we exclude financial and textual news data, averaged over 1-month and the end-of-quarter horizons.

Concluding remarks

1. Structured sparsity approach to high-dimensional time series sampled at the same or mixed frequencies
    - Legendre polynomials;
    - sparse-group LASSO.
2. Simple, scalable, and robust to model misspecification projections based on the empirical risk minimization principle.
3. Theory shows that we can handle heavy-tailed dependent data.
4. ML methods may be useful for nowcasting with high-dimensional data even when $p > T$.
5. Textual news and financial data can be a useful addition to more traditional macroeconomic data.
6. `midasml` package on CRAN (+ Matlab and Julia code on Github).

Thank you!

ababii.bitbucket.io