

Inference for Factor-MIDAS Regression Models

Julia Koh

Department of Econometrics and Operations Research
Tilburg University

January 29, 2026

Abstract

Factor-MIDAS regression models are often used to forecast a target variable using common factors extracted from a large panel of predictors observed at higher frequencies. In the paper, we derive the asymptotic distribution of the factor-MIDAS regression estimator coefficients. We show that there exists an asymptotic bias because the factors are estimated. However, the fact that factors and their lags are aggregated in a MIDAS regression model implies that the asymptotic bias depends on both serial and cross-sectional dependence in the idiosyncratic errors of the factor model. Thus, bias correction is more complicated in this setting. Our second contribution is to propose a bias correction method based on a plug-in version of the analytical formula we derive. This bias correction can be used in conjunction with asymptotic normal critical values to produce asymptotically valid inference. Alternatively, we can use a bootstrap method, which is our third contribution. We show that correcting for bias is important in simulations and in an empirical application to forecasting quarterly U.S. real GDP growth rates using monthly factors.

Keywords: Factor model, Bootstrap, Asymptotic Bias

1 Introduction

MIDAS (Mixed-Data Sampling) regressions are popular tools in forecasting. Originally proposed by Ghysels et al. (2004; 2005; 2006; 2007), these models combine predictors observed at high frequencies by relying on a parametric temporal aggregation function to forecast a target variable sampled at a lower frequency. Originally proposed to handle financial variables, they have become standard tools in macroeconomic forecasting (see e.g., Clements and Galvão (2008; 2009), which relies on MIDAS autoregressions for nowcasting U.S. real output growth).

More recently, standard MIDAS regressions have been generalized to “factor-MIDAS regressions” (or “factor-augmented MIDAS regression models”) by including as predictors common factors extracted from a large panel of time series sampled at a higher frequency than the target variable. By combining with the dimension reduction properties of factor models, factor-MIDAS regressions are powerful tools for forecasting and they are often used in empirical applications (see for instance Marcellino and Schumacher (2010), Monteforte and Moretti (2013), Kim and Swanson (2018), and Ferrara and Marsilli (2019)). Estimation of factor-MIDAS regressions is complicated by the fact that some of the predictors are latent common factors. It typically proceeds in two steps: we first extract the common factors using principal component analysis, and then estimate the model using nonlinear least squares, where the estimated factors are aggregated by a temporal aggregation scheme.

Although factor-MIDAS regressions are empirically popular, no formal inference methods have been proposed in the literature. Our paper proposes inference methods for factor-MIDAS regression models and provides the theoretical justification for these methods. The main contributions of this paper are as follows. Firstly, the asymptotic distribution of the

factor-MIDAS regression estimators is derived. We show that there is an asymptotic bias in the second step due to the estimation of the factors in the first step. Secondly, we propose two inference methods accounting for this bias: a bias correction method based on the bias formula we derive and a bootstrap method.

Our work is related to the existing literature on factor-augmented regression models (without mixed frequencies). Bai and Ng (2006) first studied the “generated regressor” problem in standard factor-augmented regression models. They showed that inference for the regression coefficients could proceed as if the estimated factors were observed if the cross-sectional dimension N was sufficiently large relative to the time dimension T , more precisely if $\sqrt{T}/N \rightarrow 0$. More recently, Gonçalves and Perron (2014) (henceforth, GP (2014)) showed that an asymptotic bias may appear under more relaxed assumption (i.e. if $\sqrt{T}/N \rightarrow c$, $0 < c < \infty$). We extend these results to factor-MIDAS regression models. This is not a trivial extension for two main reasons. First, the estimation problem in a factor-MIDAS regression model is more complicated because the predictors include latent factors (and their lags) sampled at a different frequency than a variable of interest. In addition, the second step is based on nonlinear least squares (rather than OLS) because of a temporal aggregation, and this complicates the asymptotic analysis. In particular, whereas the bias derived in Gonçalves and Perron (2014) depends only on the cross-sectional dependence, the asymptotic bias of a factor-MIDAS regression model depends on both serial and cross-sectional dependence in the idiosyncratic errors. Consequently, different methods of inference are required for factor-MIDAS regressions.

We consider two different methods of inference in this context. The first is an analytical bias correction that can be used along with asymptotic normal critical values. Our plug-in bias correction is robust to both serial and cross-sectional dependence of unknown form

in the idiosyncratic errors. It is based on the asymptotic formula of the bias we derive, replacing unknown parameters with consistent estimators. As in Ludvigson and Ng (2009), who also propose a bias correction formula for the standard factor-augmented regression model without mixed frequencies, we rely on the CS-HAC estimator of Bai and Ng (2006) to account for cross-sectional dependence. However, our estimator is more complex since it also requires robustness to serial dependence.

Our second method of inference is based on the bootstrap. The bootstrap has two significant advantages: it can perform better in finite samples, and it avoids the explicit estimation of the bias term which can be complicated in this context. We propose a bootstrap procedure inspired by Gonçalves and Perron (2014), which is a residual-based bootstrap. Although the method is inspired by Gonçalves and Perron (2014), the asymptotic justification is substantially more complicated. More importantly, the need to mimic the asymptotic bias requires the bootstrap to be robust to both serial and cross-sectional dependence. Since none of the existing bootstrap methods in the literature allows for both forms of dependence, we propose a new bootstrap method for factor models that has these properties. Our method is based on an application of the sieve bootstrap to the idiosyncratic residuals of each time series in the panel data model, where the corresponding innovations are resampled using the cross-sectional dependent bootstrap proposed by Gonçalves and Perron (2020). We show that this bootstrap method is asymptotically valid when each idiosyncratic error in the factor model is generated by an $AR(\infty)$ process with innovations that are potentially cross-sectionally correlated across the panel. A special case of this new bootstrap method is considered by Gonçalves, Koh, and Perron (2024) when testing for the number of common factors in group factor models (as proposed by Andreou, Gagliardini, Ghysels, and Rubin (2019)) without theoretical justification.

We illustrate the good finite sample performance of the plug-in bias estimator and the bootstrap using Monte Carlo simulations. In particular, the results show that it is important to correct the bias due to the estimation of the factors in the first step. Although both the plug-in bias correction and the bootstrap methods replicate the bias well, the bootstrap outperforms the plug-in bias estimator by further reducing the coverage rate distortions. Finally, we apply our new inference methods to an empirical application where we nowcast quarterly U.S. real GDP growth rate using monthly macroeconomic factors. The results show that there is a significant bias, thereby indicating the importance of correcting it.

The rest of this paper is organized as follows. In Section 2, we derive the asymptotic distribution of the factor-augmented MIDAS regression model and propose a plug-in bias estimator. In Section 3, we propose and theoretically justify the bootstrap. The simulation results are shown in Section 4, and the empirical application is discussed in Section 5. Section 6 concludes the paper.

For any matrix A , $\|A\|$ denotes its Frobenius norm defined as $\|A\| = (\text{trace}(A'A))^{1/2}$. $\rho(A)$ denotes the Euclidean vector norm of the vector Ax : $\rho(A) = \max_{\|x\|=1} \|Ax\|$, where $\|Ax\| = (x'A'Ax)^{1/2}$.

2 Asymptotic Theory

2.1 Factor-augmented MIDAS regression models

The MIDAS regression model projects high-frequency variables onto a target variable, which is denoted as y_t . The regressors are observed at most m times between t and $t-1$. To handle variables sampled at mixed frequency, a MIDAS regression aggregates the high-frequency variables with a lag polynomial function. The basic MIDAS regression model with a single

observed regressor x_t can be written as follows:

$$y_t = \beta_0 + \beta_1 W(L^{1/m}; \theta) x_t + \varepsilon_t, \quad t = 1, \dots, T, \quad (1)$$

where $W(L^{1/m}; \theta) = \sum_{k=1}^K w_k(\theta) L^{k/m}$ and $L^{k/m} x_t = x_{t-k/m}$. Here, $w_k(\theta)$ is a weighting function that temporally aggregates the regressor and its lags, and θ is a $p \times 1$ vector of weighting parameters. To identify β_1 , we assume that $w_k(\theta) \in (0, 1)$ and $\sum_{k=1}^K w_k(\theta) = 1$. A common weighting scheme in the MIDAS regression model is the exponential Almon lag with two parameters such that

$$w_k(\theta) = \frac{\exp(\theta_1 k + \theta_2 k^2)}{\sum_{k=1}^K \exp(\theta_1 k + \theta_2 k^2)}. \quad (2)$$

Other weighting schemes include the beta function and the linear function (see Ghysels, Valkanov, and Serrano (2009) for detail).

In this paper, we consider the factor-MIDAS regression model, which employs unobserved high-frequency factors as regressors. In particular, letting the regressor x_t in (1) be replaced by a latent factor, we write the model as follows.

$$y_t = \beta_0 + \beta_1 W(L^{1/m}; \theta) f_t + \varepsilon_t = \beta_0 + \beta_1 \sum_{k=1}^K w_k(\theta) f_{t-k/m} + \varepsilon_t, \quad t = 1, \dots, T,$$

where $f_{t-k/m}$ is a (single) factor in the following panel factor model,

$$X_{t-k/m} = \Lambda f_{t-k/m} + e_{t-k/m}, \quad k = m-1, \dots, 0, \quad \text{and } t = 1, \dots, T. \quad (3)$$

The factor model includes factor loadings denoted by Λ and an idiosyncratic error term, $e_{t-k/m}$. If there are r unobserved factors, represented by a $r \times 1$ vector of common factors

denoted by $f_{t-k/m}$ in the factor model (3), then the model can be generalized as follows.

$$y_t = \beta_0 + \beta_1' W(L^{1/m}; \theta) f_t + \varepsilon_t = \beta_0 + \beta_1' F_t(\theta) + \varepsilon_t, \quad t = 1, \dots, T, \quad (4)$$

where $\beta_1 = (\beta_{1,1}, \dots, \beta_{1,r})'$, and $\theta = (\theta_1', \dots, \theta_r')$ with $\theta_j = (\theta_{j,1}, \dots, \theta_{j,p})'$, a $p \times 1$ weighting parameter¹ for j -th factor, for $j = 1, \dots, r$. We define $F_t(\theta) \equiv W(L^{1/m}; \theta) f_t$ in the second equality. In fact, the temporal aggregation in this generalized model applies on a vector as

$$F_t(\theta) = \sum_{k=1}^K w_k(\theta) L^{k/m} f_t = \sum_{k=1}^K w_k(\theta) f_{t-k/m},$$

where $w_k(\theta)$ is a $r \times r$ diagonal matrix such that $w_k(\theta) \equiv \text{diag}(w_{k,1}(\theta_1), \dots, w_{k,r}(\theta_r))$, where $w_{k,j}(\theta_j)$ is the weight for the k -th lag of the j -th factor.² To derive the distribution in the next section, we further simplify the general factor-MIDAS regression model (4) to

$$y_t = g(F_t, \alpha) + \varepsilon_t, \quad t = 1, \dots, T, \quad (5)$$

where $g(F_t, \alpha) = \beta_0 + \beta_1' F_t(\theta)$, $\alpha = (\beta', \theta')$ with $\beta = (\beta_0, \beta_1')$, and $F_t = (1, f_t', f_{t-1/m}', \dots, f_{t-K/m}')'$.

For convenience, we use the high frequency time index denoted by $t_h = 1, \dots, T_H$, where $T_H = mT$. We derive this by noting that $t_h = m((t-1) + i/m)$ for $i = 1, \dots, m$, and $t = 1, \dots, T$.³ Using this notation, we can write the factor model as $X_{t_h} = \Lambda f_{t_h} + e_{t_h}$, for $t_h = 1, \dots, T_H$. Using the matrix notation, we write the factor model as $X = f\Lambda' + e$, where X is a $T_H \times N$ matrix of high-frequency time series, $f = (f_1, \dots, f_{T_H})'$ is a $T_H \times r$ matrix of common factors, and e is a $T_H \times N$ matrix of idiosyncratic errors.⁴

¹Note that at least one component of β_1 needs to be non-zero to identify the weighting parameters, θ .

²Note that when $m = 1$ and $K = 0$, the factor-MIDAS regression model is equivalent to the standard factor-augmented regression model in GP (2014).

³With this notation, a high-frequency observation at t_h is equivalent to observing it at the i -th intra-period between $t-1$ and t . Note that the time notation in the factor model (3) can be written as $(t-1) + (m-k)/m$.

⁴One may consider a situation where X includes variables with different frequencies, such as monthly and quarterly, while y_t is observed annually. In this case, the group factor model discussed in Andreou et al. (2019) can be exploited to extract the factors.

2.2 Asymptotic Theory

We denote NLS estimators by $\hat{\alpha}$ when the factors are observed. Then, Andreou, Ghysels, and Kourtellis (2010) show that the limiting distribution of $\hat{\alpha}$ is as following:

$$\sqrt{T}(\hat{\alpha} - \alpha_0) \xrightarrow{d} N(0, \Sigma^{-1}\Omega\Sigma^{-1}), \quad (6)$$

where $\alpha_0 = (\beta', \theta)'$, $\Sigma = E[g_{\alpha,t}g'_{\alpha,t}]$, and $\Omega = E[\varepsilon_t^2 g_{\alpha,t}g'_{\alpha,t}]$ with $g_{\alpha,t} = \partial g(F_t, \alpha)/\partial \alpha$. When the true factors are observed, the estimators are normally distributed with mean zero and a sandwich variance.

In factor-MIDAS models, however, the factors are latent, and we have to estimate them. Accordingly, the estimation in the factor-MIDAS regression model proceeds in two steps. First, we estimate the common factors from a panel dataset of high-frequency indicators by principal component analysis (PCA). The estimated factors, \tilde{f} , are equivalent to $\sqrt{T_H}$ times the eigenvectors of $XX'/T_H N$ corresponding to the r largest eigenvalues (in decreasing order). The estimated factor loadings are $\tilde{\Lambda} = X'\tilde{f}/T_H$.⁵ Second, we estimate the parameters β and θ using nonlinear least squares (NLS) by regressing the low frequency variable on the temporally aggregated estimated factors at high-frequency. In the factor model, the estimated factors \tilde{f}_t are only consistent for Hf_t , where the rotation matrix H is defined as $H = \tilde{V}^{-1}\frac{\tilde{f}'f}{T_H}\frac{\Lambda'\Lambda}{N}$, and \tilde{V} is a $r \times r$ diagonal matrix of eigenvalues of $XX'/T_H N$ in a descending order (for more details, see Bai (2003)). By incorporating the estimated factors in the regression and noting the rotation of the factors, we can rewrite (4) as follows.

$$y_t = \beta_0 + \beta_1' H^{-1} \tilde{F}_t(\theta) + \beta_1' H^{-1} (H F_t(\theta) - \tilde{F}_t(\theta)) + \varepsilon_t = g(\tilde{F}_t, \alpha) + \xi_t, \quad (7)$$

⁵When $T_H > N$, we use normalization such that $\Lambda'\Lambda/N = I_r$ and $f'f$ is a diagonal matrix, which is computationally easier. In this case, $\tilde{\Lambda}$ is the matrix of \sqrt{N} times the eigenvectors of $XX'/T_H N$ corresponding to the r largest eigenvalues and the estimated factors are $\tilde{f} = X\tilde{\Lambda}/N$.

where $g(\tilde{F}_t, \alpha) = \beta_0 + \beta_1' H^{-1} \tilde{F}_t(\theta)$, $\alpha = (\beta_0, \beta_1' H^{-1}, \theta')'$, and $\tilde{F}_t(\theta) = \sum_{k=1}^K w_k(\theta) \tilde{f}_{t-k/m}$. The coefficient on the aggregated factors estimates $\beta_1' H^{-1}$. Moreover, the estimation error of the factors implies that the regression error term is $\xi_t = \beta_1' H^{-1} (H F_t(\theta) - \tilde{F}_t(\theta)) + \varepsilon_t$. We denote the NLS estimators of α in (7) by $\tilde{\alpha} = (\tilde{\beta}', \tilde{\theta}')'$ to distinguish from $\hat{\alpha} = (\hat{\beta}', \hat{\theta}')'$, which are the estimators from the regression of y_t on the true factors f_t . Next, we derive the limiting distribution of $\sqrt{T}(\tilde{\alpha} - \alpha)$ under the assumption that $\sqrt{T}/N \rightarrow c$, where $0 \leq c < \infty$. Note that although the variable of interest is a linear function of factor estimation error similar to the factor-augmented regression models, there exists a nonlinear weighting function. Furthermore, unlike standard factor-augmented regression models, the lags of the factors are incorporated. As will be demonstrated in the next theorem, the incorporation of the lags of the factors results in the fact that the asymptotic bias relies on the time-series dependence and cross-sectional dependence in the idiosyncratic error term.⁶

The asymptotic distribution of the estimators is derived under Assumptions A.1 - A.6 in Section A in Online Appendix. We also introduce the following notations: $V \equiv \text{plim } \tilde{V}$, $Q \equiv \text{plim} \left(\frac{\tilde{f}' f}{T_H} \right)$, $Q_k \equiv \text{plim} \left(\frac{1}{T_H - k} \sum_{t_h=k+1}^{T_H} \tilde{f}'_{t_h} f_{t_h-k} \right)$, and $\Sigma_{\tilde{f}} \equiv V^{-1} Q \Gamma Q' V^{-1}$, which is the asymptotic variance of $\sqrt{N}(\tilde{f}_{t_h} - H f_{t_h})$.⁷ The asymptotic variance of the factor estimation error is a function of Γ , which is defined by $\Gamma \equiv \lim_{N \rightarrow \infty} \text{Var} \left(\frac{\Lambda' e_{t_h}}{\sqrt{N}} \right)$. We assume that the idiosyncratic errors in the factor model, e_{t_h} is stationary in Assumption A.2-(d). Under the stationarity of the idiosyncratic errors, we also denote $\Gamma_k \equiv \lim_{N \rightarrow \infty} \text{Cov} \left(\frac{\Lambda' e_{t_h-k}}{\sqrt{N}}, \frac{\Lambda' e_{t_h}}{\sqrt{N}} \right)$. Note that by the identification assumption, Assumption A.1-(d) in Online Appendix, we have $Q = H_0$, where $H_0 = \text{plim } H$, and H_0 is a diagonal matrix of ± 1 , where the sign is determined by the sign of $\tilde{f}' f / T_H$ (for the detail of the proof, see the proof of (2) in Bai and

⁶Note that the time-series dependence in the idiosyncratic error term does not appear in the asymptotic bias in the standard factor augmented regression models. For detail, see GP (2014) (their Theorem 2.1).

⁷For the details, see Bai (2003).

Ng (2013)). Therefore, the asymptotic variance can be also written as $\Sigma_{\tilde{f}} = V^{-1}H_0\Gamma H'_0V^{-1}$.

Theorem 2.1 (Asymptotic distribution of the estimators in the factor-MIDAS models)

If $\sqrt{T}/N \rightarrow c$, where $0 \leq c < \infty$, and Assumptions A.1 - A.6 in Section A in Online Appendix hold,

$$\sqrt{T}(\tilde{\alpha} - \alpha) \xrightarrow{d} N(-c\Delta_\alpha, \Sigma_\alpha), \quad (8)$$

where $\Sigma_\alpha \equiv \Phi_0'^{-1}\Sigma^{-1}\Omega\Sigma^{-1}\Phi_0^{-1}$ with $\Phi_0 = \text{diag}(1, H_0, I_p)$, and

$$\Delta_\alpha = \begin{bmatrix} \Delta_\beta \\ \Delta_\theta \end{bmatrix} = (\Phi_0\Sigma\Phi_0')^{-1} \begin{bmatrix} B_\beta \\ B_\theta \end{bmatrix}. \quad (9)$$

$B_\beta = (B_{\beta_0}, B'_{\beta_1})'$ and B_θ are such that $B_{\beta_0} = 0$,

$$\begin{aligned} B_{\beta_1} &= \left[\sum_{k=1}^K w_k(\theta) \left\{ \Sigma_{\tilde{f}} + V\Sigma_{\tilde{f}}V^{-1} \right\} w_k(\theta) \right. \\ &\quad \left. + \sum_{k=1}^K \sum_{l \neq k}^K w_k(\theta) \left\{ V^{-1}H_0\Gamma_{k-l}H'_0V^{-1} + Q_{k-l}\Gamma H'_0V^{-2} \right\} w_l(\theta) \right] \text{plim}(\tilde{\beta}_1), \end{aligned} \quad (10)$$

and

$$\begin{aligned} B_\theta &= \text{plim}(\tilde{\beta}_1) \circ \left[\sum_{k=1}^K \frac{\partial w_k(\theta)}{\partial \theta} \left\{ \Sigma_{\tilde{f}} + V\Sigma_{\tilde{f}}V^{-1} \right\} w_k(\theta) \right. \\ &\quad \left. + \sum_{k=1}^K \sum_{l \neq k}^K \frac{\partial w_k(\theta)}{\partial \theta} \left\{ V^{-1}H_0\Gamma_{k-l}H'_0V^{-1} + Q_{k-l}\Gamma H'_0V^{-2} \right\} w_l(\theta) \right] \text{plim}(\tilde{\beta}_1), \end{aligned} \quad (11)$$

where $\frac{\partial w_k(\theta)}{\partial \theta} \equiv \text{diag} \left(\frac{\partial w_{k,1}(\theta_1)}{\partial \theta_1}, \dots, \frac{\partial w_{k,r}(\theta_r)}{\partial \theta_r} \right)$ is a block diagonal matrix and the j -th diagonal block is a $p \times 1$ vector given by $\frac{\partial w_{k,j}(\theta_j)}{\partial \theta_j}$ for $j = 1, \dots, r$.

In (11) in Theorem 2.1, we use the Hadamard product which is equivalent to $(A \circ B)_{ij} = A_{ij}B_{ij}$. More specifically, $\beta \circ \frac{\partial w_k(\theta)}{\partial \theta}$ is a block diagonal matrix where the j -th diagonal block contains $\beta_j \frac{\partial w_{j,k}(\theta_j)}{\partial \theta_j}$ for $j = 1, \dots, r$. Based on Theorem 2.1, the bias of the estimators is

proportional to c , the limiting value of \sqrt{T}/N , and also to $\text{plim}(\tilde{\beta}_1) = (H_0^{-1})'\beta_1$. This implies that the estimates are biased unless $\beta_1 = 0$ or $c = 0$. Additionally, the asymptotic variance of the estimated factors, $\Sigma_{\tilde{f}}$, affects the bias. Since the variance of the factor estimation error depends on Γ , which is a variance of the scaled average of the factor loadings and the idiosyncratic errors in the factor model, the cross-sectional dependence of factor errors matters. These findings are similar to the bias in the context of GP (2014).

It is important to highlight two main differences in the asymptotic bias between the factor-MIDAS regression model and standard factor-augmented regression models. Firstly, the bias in the MIDAS regression model depends on the weighting scheme, $w_k(\theta)$, due to a temporal aggregation.⁸ Secondly, the bias depends on the covariance of the cross-sectional average of factor loadings and the idiosyncratic error terms between two distinct periods, represented as Γ_{k-l} . This term arises due to the presence of the lags of the estimated factors. To see this, consider a simple factor-augmented regression model with a lag and without mixed-frequency variables as follows.

$$y_t = \beta_1 f_t + \beta_2 f_{t-1} + \varepsilon_t = \beta' F_t + \varepsilon_t,$$

where $\beta = (\beta_1, \beta_2)'$ and $F_t = (f_t, f_{t-1})'$. We assume that the factor is a single factor for simplicity. By the fact that the factors are estimated, we can rewrite it as follows.

$$y_t = \beta' H^{-1} \tilde{F}_t + \beta' H^{-1} (H F_t - \tilde{F}_t) + \varepsilon_t.$$

Note that since we include a lag of the factor, we have a factor estimation error at $t - 1$ as well as contemporaneous factor estimation error. Letting $\hat{\beta}$ be OLS estimator from a

⁸When there is no temporal aggregation, the MIDAS regression becomes unrestricted MIDAS (U-MIDAS) proposed by Foroni, Marcellino, and Schumacher (2015). If the estimated factors are used as predictors in U-MIDAS, there will be bias that depends on cross-sectional and serial dependence of the idiosyncratic error term in the factor model, by the fact that lags of the estimated factors are present.

regression of y_t on \tilde{F}_t , we can show that

$$\sqrt{T}(\hat{\beta} - H^{-1}\beta) = \left(\frac{1}{T} \sum_{t=1}^T \tilde{F}_t \tilde{F}_t' \right)^{-1} \frac{1}{\sqrt{T}} \sum_{t=1}^T \tilde{F}_t \varepsilon_t + \sqrt{T} \left(\frac{1}{T} \sum_{t=1}^T \tilde{F}_t \tilde{F}_t' \right)^{-1} \frac{1}{T} \sum_{t=1}^T \tilde{F}_t (HF_t - \tilde{F}_t)' H^{-1} \beta.$$

In fact, we can show that $\frac{1}{T} \sum_{t=1}^T \tilde{F}_t (\tilde{F}_t - HF_t)' \xrightarrow{p} \frac{1}{N} \left(\frac{1}{T} \sum_{t=1}^T \text{Var} \left(\sqrt{N}(\tilde{F}_t - HF_t) \right) \right) = O_p(1/N)$ by Bai (2003) (see their Lemma B.2) and GP (2014). Therefore, the second term is $O_p(\sqrt{T}/N)$ and drives the asymptotic bias under the rate condition such that $\sqrt{T}/N \rightarrow c$ for $0 \leq c < \infty$. In GP (2014), since the contemporaneous factor is the sole predictor in their factor-augmented regression model, the variance of contemporaneous factor estimation error appears alone. More specifically, the bias is driven by $\frac{1}{T} \sum_{t=1}^T \text{Var} \left(\sqrt{N}(\tilde{f}_t - Hf_t) \right)$, which depends on $\frac{1}{T} \sum_{t=1}^T \text{Var} \left(\frac{\Lambda' e_t}{\sqrt{N}} \right)$. This term implies that the bias depends solely on the cross-sectional dependence of the idiosyncratic error term in the factor model. However, when we incorporate a lag of the factor as a predictor alongside the contemporaneous factor, the covariance between the factor estimation error at t and $t - 1$ becomes relevant, which depends on $\frac{1}{T} \sum_{t=1}^T \text{Cov} \left(\frac{\Lambda' e_t}{\sqrt{N}}, \frac{\Lambda' e_{t-1}}{\sqrt{N}} \right)$. Thus, the inclusion of the lag of the factor indicates that the bias depends not only on the cross-sectional dependence, but also on the time-series dependence of the idiosyncratic error term in the factor model.⁹

In the factor-MIDAS regression model, the inclusion of lagged estimated factors introduces additional complexity. Similar to the previously discussed simple case, we have an extra term such that $\frac{1}{T_{H-k}} \sum_{t_h=k+1}^{T_H} \text{Cov}(\sqrt{N}(\tilde{f}_{t_h} - Hf_{t_h}), \sqrt{N}(\tilde{f}_{t_h-k} - Hf_{t_h-k}))$ for $k \neq 0$, which depends on $\frac{1}{T_{H-k}} \sum_{t_h=k+1}^{T_H} \text{Cov} \left(\frac{\Lambda' e_{t_h}}{\sqrt{N}}, \frac{\Lambda' e_{t_h-k}}{\sqrt{N}} \right)$. Therefore, the bias in our context relies on the serial dependence as well as cross-sectional dependence of the idiosyncratic error term in the factor model. This finding holds considerable significance, as the literature surround-

⁹This also explains why the bias in unrestricted MIDAS (U-MIDAS) regression models augmented by the factors depends on cross-sectional as well as serial dependence of the idiosyncratic error term in the factor model.

ing factor-augmented regression models has primarily concentrated on the cross-sectional dependence of the idiosyncratic error term. This focus necessitates the development of novel inference methods that can effectively account for the time-series dependence inherent in the idiosyncratic error term, which appears in our context.

2.3 Plug-in Bias

In this section, we propose an analytical estimator to account for the bias identified in Theorem 2.1. This is inspired by Ludvigson and Ng (2009), where they propose a plug-in bias estimator by replacing the unknown quantities with their consistent estimators and correcting the bias in the context of the factor-augmented regression model. Similarly, we propose a bias-corrected estimator for factor-augmented MIDAS regression models.

In order to do that, we need a consistent estimator for the term Γ_k , which has never been explored previously. Note that it depends on the cross-sectional and the serial dependence of the idiosyncratic error term. When the idiosyncratic error term is serially but not cross-sectionally correlated, we can estimate this term as $\hat{\Gamma}_k = \frac{1}{N(T_H-k)} \sum_{t_h=k+1}^{T_H} \sum_{i=1}^N \tilde{\lambda}_i \tilde{\lambda}_i' \tilde{e}_{i,t_h} \tilde{e}_{i,t_h-k}$, where $\hat{\Gamma}_k$ denotes the estimator of Γ_k . However, when the idiosyncratic error term is cross-sectionally and serially dependent, estimating this term is no longer straightforward, as discussed in Bai and Ng (2006). To address this issue, Bai and Ng (2006) propose an estimator for the variance-covariance matrix of the cross-sectional average of factor loadings and the idiosyncratic error term, denoted by Γ . They use the time series observations and truncation with $n < N$ under the covariance stationarity such that $\hat{\Gamma}_{\text{CS-HAC}} = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \tilde{\lambda}_i \tilde{\lambda}_j' \frac{1}{T_H} \sum_{t_h=1}^{T_H} \tilde{e}_{i,t_h} \tilde{e}_{j,t_h}$.

To propose a method to estimate Γ_k that takes into account cross-sectional and serial dependence, we take an approach, similar to the one used in Bai and Ng (2006). We use

the time series observations and a truncation method, that limits $n < N$ observations. We denote the estimator for Γ_k by $\hat{\Gamma}_k$, which is defined as follows.

$$\hat{\Gamma}_{k,\text{CS-HAC}} = \frac{1}{T_H - k} \sum_{t_h=k+1}^{T_H} \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \tilde{\lambda}_i \tilde{\lambda}'_j \tilde{e}_{i,t_h} \tilde{e}_{j,t_h-k}, \quad (12)$$

where $n = \min(\sqrt{N}, \sqrt{T_H})$. Note that by Assumption A.2-(d), Γ_k does not depend on time.

Theorem 2.2 *Suppose the Assumptions A.1 - A.4 in Section A in Online Appendix hold.*

Then, for any fixed $k = 0, 1, 2, \dots, K - 1$

$$\|\hat{\Gamma}_k - H_0^{-1'} \Gamma_k H_0^{-1}\| \xrightarrow{p} 0 \quad \text{if} \quad \frac{n}{\min(N, T_H)} \rightarrow 0,$$

Here, in Theorem 2.2, $\hat{\Gamma}_k$ depends on the assumption on the serial and cross-sectional dependence in the idiosyncratic errors of the factor model. If there is only serial dependence, $\hat{\Gamma}_k = \frac{1}{N} \sum_{i=1}^N \tilde{\lambda}_i \tilde{\lambda}'_i \frac{1}{T_H - k} \sum_{t_h=k+1}^{T_H} \tilde{e}_{i,t_h} \tilde{e}_{i,t_h-k}$. If we allow for cross-sectionally dependence additionally, $\hat{\Gamma}_k = \tilde{\Gamma}_{k,\text{CS-HAC}}$ defined in (12). Note that if $k = 0$, our estimators are equivalent to the estimators proposed in Bai and Ng (2006). Theorem 2.2 enables us to construct consistent estimators for (10) and (11) as follows.

$$\begin{aligned} \hat{B}_{\beta_1} &= \left[2 \sum_{k=1}^K w_k(\tilde{\theta}) \tilde{\Sigma}_{\tilde{f}} w_k(\tilde{\theta}) + \sum_{k=1}^K \sum_{l \neq k}^K w_k(\tilde{\theta}) \left\{ \tilde{V}^{-1} \hat{\Gamma}_{k-l,\text{CS-HAC}} \tilde{V}'^{-1} + \tilde{Q}_{k-l} \hat{\Gamma}_{\text{CS-HAC}} \tilde{V}^{-2} \right\} w_l(\tilde{\theta}) \right] \tilde{\beta}_1, \text{ and} \\ \hat{B}_{\theta} &= \tilde{\beta}_1 \circ \left[2 \sum_{k=1}^K \frac{\partial w_k(\tilde{\theta})}{\partial \theta} \tilde{\Sigma}_{\tilde{f}} w_k(\tilde{\theta}) + \sum_{k=1}^K \sum_{l \neq k}^K \frac{\partial w_k(\tilde{\theta})}{\partial \theta} \left\{ \tilde{V}^{-1} \hat{\Gamma}_{k-l,\text{CS-HAC}} \tilde{V}^{-1} + \tilde{Q}_{k-l} \hat{\Gamma}_{\text{CS-HAC}} \tilde{V}^{-2} \right\} w_l(\tilde{\theta}) \right] \tilde{\beta}_1, \end{aligned}$$

where $\tilde{\Sigma}_{\tilde{f}} = \tilde{V}^{-1} \tilde{Q} \hat{\Gamma}_{\text{CS-HAC}} \tilde{Q} \tilde{V}^{-1}$ with $\tilde{Q} = \tilde{f}' \tilde{f} / T_H$, and $\tilde{Q}_{k-l} = \sum_{t_h=k+1}^{T_H} \tilde{f}'_{t_h} \tilde{f}_{t_h-k}$. Note that the bias estimates can be simpler under the restriction on either cross-sectional or serial dependence, or both. We denote the bias-corrected estimator by $\hat{\alpha}_{\text{BC}}$ such that $\hat{\alpha}_{\text{BC}} \equiv \tilde{\alpha} - (-\frac{1}{N} \hat{\Delta}_{\alpha})$. Here, $-\hat{\Delta}_{\alpha}$ is the estimate of the bias in $\tilde{\alpha}$, where $\hat{\Delta}_{\alpha} = \hat{\Sigma}^{-1} (\hat{B}'_{\beta}, \hat{B}'_{\theta})'$ with $\hat{\Sigma}$ a consistent estimator of Σ , $\hat{B}_{\beta} = (\hat{B}_{\beta_0}, \hat{B}_{\beta_1})'$, and $\hat{B}_{\beta_0} = 0$.

Proposition 2.1 *Suppose the Assumptions A.1 - A.6 in Section A in Online Appendix hold and $\sqrt{T}/N \rightarrow c$, where $0 \leq c < \infty$, then*

$$\sqrt{T}(\hat{\alpha}_{BC} - \alpha) \xrightarrow{d} N(0, \Sigma_\alpha). \quad (13)$$

Based on Proposition 2.1, the bias corrected estimator no longer contains an asymptotic bias. However, it is well known that an approach based on asymptotic theory does not perform well in finite samples. Additionally, the bias takes a very complicated form in our context, which makes it difficult to implement. Therefore, we discuss an alternative approach, a bootstrap method in the next section.

3 Bootstrap method: AR-sieve+CSD bootstrap

In this section, we propose a bootstrap method and show its validity by proving that our method satisfies bootstrap high level conditions under which any general residual-based bootstrap is satisfied. We leave the bootstrap high level conditions in the appendix (see Section C in the Online Appendix).

In particular, we propose a bootstrap procedure, where we resample the factor model and the MIDAS regression model, and then obtain the bootstrap estimates. Note that in Theorem 2.1, we show that the asymptotic bias in our context relies on the cross-sectional and serial dependence in the idiosyncratic error term in the factor model, therefore, it is crucial that the bootstrap resampled idiosyncratic error term in the factor model mimics these dependences. To the best of our knowledge, replicating the time-series dependence in the error term in the factor model has not been studied in the literature. GP (2014) propose a wild bootstrap and prove its validity in the context of the factor-augmented regression

models under no cross-sectional dependence in the error term in the factor model.¹⁰ To allow for cross-sectional dependence, Gonçalves and Perron (2020) propose a bootstrap method that utilizes a thresholding technique to allow for the cross-sectional dependence, so-called CSD (cross-sectional dependent) bootstrap. However, these methods cannot be used in our context as it destroys the serial dependence in the idiosyncratic error terms.

On the other hand, to resample the error term in the MIDAS regression model, GP (2014) propose a wild bootstrap under the assumption that the regression error terms follow martingale difference sequence. Djogbenou, Gonçalves, and Perron (2015) propose a block wild bootstrap and a dependent wild bootstrap to resample the regression error terms to account for serially correlated regression error terms. Depending on the assumption a researcher is willing to make, either the approach proposed by GP (2014) or by Djogbenou et al. (2015) can be similarly applied to resample the regression error terms in our context. In this paper, for simplicity, we rely on the assumption that the regression error terms follow martingale difference sequence and use the wild bootstrap.

The key finding in our paper is that the bias within our framework is influenced by both serial and cross-sectional dependence in the idiosyncratic error term in the factor model. To address this, we propose a novel bootstrap method that can replicate both dependences. Specifically, we combine autoregressive sieve bootstrap and the CSD bootstrap to resample the residuals in the factor model.¹¹ The autoregressive sieve bootstrap, initially introduced by Bühlmann (1997) and further explored by Kreiss, Paparoditis, and Politis (2011) and Meyer and Kreiss (2015), has been effectively applied to the estimated factors by Bi, Shang, Yang,

¹⁰Note that the asymptotic bias in the factor augmented regression models studied in GP (2014) only depends on the cross-sectional dependence. For detail, see GP (2014).

¹¹Note that we cannot use block-based bootstrap or dependent wild bootstrap to account for serial dependence, because these bootstrap methods induce a zero cross-sectional dependence. (For detail, see Gonçalves and Perron (2020).)

and Zhu (2021). In our paper, we combine this method with the CSD bootstrap method and apply it to the residuals in the factor model, which we refer to as the AR-sieve+CSD bootstrap method. A more restricted version of our approach is recently considered by Gonçalves et al. (2024), where they substitute the autoregressive sieve bootstrap with an autoregressive parametric bootstrap of order one, albeit without theoretical justification. Also, as addressed in Bühlmann (1997), the autoregressive sieve bootstrap method offers more flexibility than a parametric autoregressive model, which is highly subject to model misspecification. The AR-sieve+CSD bootstrap method resamples each time series residual in the factor model through an autoregressive sieve process, while the corresponding innovations are resampled by the CSD bootstrap method. This approach effectively captures cross-sectional dependence in the innovation terms through the CSD bootstrap method and the serial dependence through the autoregressive process. The detailed algorithm to use the AR-sieve+CSD bootstrap to resample the residuals in the factor model can be found in Algorithm 1.¹² In Algorithm 1, we resample the residuals in the factor model similar to the bootstrap procedure in Kreiss et al. (2011) and Bühlmann (1997). The difference is that we resample the innovation terms in the autoregressive process using CSD bootstrap proposed by Gonçalves and Perron (2020).

One might consider utilizing high-dimensional vector autoregressive (VAR) models to resample the idiosyncratic error term in the factor model. Recent studies, such as those by Kock and Callot (2015) and Krampe, Kreiss, and Paparoditis (2021), have explored this high-dimensional VAR model. Kock and Callot (2015) establishes oracle inequalities for both LASSO and adaptive LASSO estimators in the context of high-dimensional VAR models. Meanwhile, Krampe et al. (2021) develops a bootstrap method applicable to this framework.

¹²The full bootstrap procedure to obtain the bootstrap estimators can be found in the Online Appendix.

Algorithm 1 : AR-sieve + CSD Bootstrap for the factor model

For $t_h = 1, \dots, T_H$, let

$$X_{i,t_h}^* = \tilde{\lambda}'_i \tilde{f}_{t_h} + e_{i,t_h}^* \quad \text{and} \quad X_{t_h}^* = \tilde{\Lambda} \tilde{f}_{t_h} + e_{t_h}^*,$$

where e_{i,t_h}^* is obtained as follows.

For each $i = 1, \dots, N$, select an order $p_i = p_i(T_H)$, $p_i \ll T_H$, for example, by an information criterion such as the Akaike information criterion (AIC), and fit a p_i -th order autoregressive model to $\tilde{e}_{i,1}, \dots, \tilde{e}_{i,T_H}$, where $\tilde{e}_{i,t_h} = X_{i,t_h} - \tilde{\lambda}'_i \tilde{f}_{t_h}$. We denote $\tilde{\phi}_i(p_i) = (\tilde{\phi}_{i,j}(p_i), j = 1, \dots, p_i)$, the Yule-Walker autoregressive parameter estimators, such that $\tilde{\phi}_i(p_i) = \tilde{\Gamma}(p_i)^{-1} \tilde{\gamma}_{p_i}$, with $\tilde{\gamma}_{p_i} = (\tilde{\gamma}_e(1), \tilde{\gamma}_e(2), \dots, \tilde{\gamma}_e(p_i))'$ and $\tilde{\Gamma}(p_i) = (\tilde{\gamma}_e(r-s))_{r,s=1,2,\dots,p_i}$ such that

$$\tilde{\gamma}_e(\tau) = \frac{1}{T_H} \sum_{t_h=1}^{T_H-|\tau|} (\tilde{e}_{i,t_h} - \bar{e}_i)(\tilde{e}_{i,t_h+|\tau|} - \bar{e}_i), \quad (14)$$

for $\tau = 0, \dots, p_i$ and $\bar{e}_i = T_H^{-1} \sum_{t_h=1}^{T_H} \tilde{e}_{i,t_h}$.

With chosen lag length $p_i = p_i(T_H)$,

$$e_{i,t_h}^* = \sum_{j=1}^{p_i} \tilde{\phi}_{i,j}(p_i) e_{i,t_h-j}^* + u_{i,t_h}^*, \quad \text{for } t_h = 1, \dots, T_H, \quad (15)$$

where $u_{t_h}^* = (u_{1,t_h}^*, \dots, u_{N,t_h}^*) = \tilde{\Sigma}_u^{1/2} \eta_{t_h}$ with $\eta_{t_h} \sim \text{i.i.d.}(0, I_N)$. The initial conditions are $e_{i,0}^*, \dots, e_{i,1-p_i}^* = 0$, for $i = 1, \dots, N$, which is equivalent to the stationary mean of e_{i,t_h}^* in the bootstrap world. Following Gonçalves and Perron (2020), we choose $\tilde{\Sigma}_u$ by a thresholding technique such that

$$\tilde{\Sigma}_u = (\hat{\sigma}_{u,ij})_{i,j=1,\dots,N},$$

with

$$\hat{\sigma}_{u,ij} = \begin{cases} \tilde{\sigma}_{u,ij} & i = j \\ \tilde{\sigma}_{u,ij} 1(|\tilde{\sigma}_{u,ij}| > \omega) & i \neq j, \end{cases} \quad \text{with } \tilde{\sigma}_{u,ij} = \frac{1}{T_H} \sum_{t_h=1}^{T_H} \tilde{u}_{i,t_h} \tilde{u}_{j,t_h},$$

where ω is a threshold and $\tilde{u}_{i,t_h} = \tilde{e}_{i,t_h} - \sum_{j=1}^{p_i} \tilde{\phi}_{i,j}(p_i) \tilde{e}_{i,t_h-j}$ for $i = 1, \dots, N$ and $t_h = 1 + p_i, \dots, T_H$.

In our paper, we do not address the high-dimensional VAR model due to the complexities involved in its theoretical justification in our framework, opting instead to reserve this for future research.

In order to prove our bootstrap method is valid, we assume that $\{e_{i,t_h}\}_{t_h=1}^{T_H}$ is an infi-

nite order moving average process that can be represented as an $\text{AR}(\infty)$ process such that $e_{i,t_h} = \sum_{j=1}^{\infty} \phi_{i,j} e_{i,t_h-j} + u_{i,t_h}$, for $t_h = 1, \dots, T_H$ and $i = 1, \dots, N$. The innovation terms in $\text{AR}(\infty)$ process, $u_{t_h} = (u_{1,t_h}, \dots, u_{N,t_h})'$, are identically and independently distributed from a distribution with mean zero and finite variance, Σ_u . Here, Σ_u is assumed to be non-diagonal to account for cross-sectional dependence in the idiosyncratic error term. More formal representation of the assumptions on our bootstrap method is provided below.

Assumption 1 λ_i are either deterministic such that $\|\lambda_i\| \leq M \leq \infty$, or stochastic such that $E\|\lambda_i\|^{24} \leq M < \infty$ for all i : $E\|f_{t_h}\|^{24} \leq M < \infty$; $E|e_{i,t_h}|^{24} \leq M < \infty$, for all (i, t_h) ; and for some $q > 1$, $E|\varepsilon_t|^{4q} \leq M < \infty$, for all t .

Assumption 2 $E(\varepsilon_t | y_t, F_t, y_{t-1}, F_{t-1}, \dots) = 0$, and $F_t = (f_{t-1/m}, \dots, f_{t-k/m})'$ and ε_t are independent of the idiosyncratic errors e_{i,s_h} for all (i, s_h, t) .

Assumption 3 $e_{i,t_h} = \sum_{j=1}^{\infty} \phi_{i,j} e_{i,t_h-j} + u_{i,t_h}$, with $\sum_{j=1}^{\infty} (1+|j|)^r |\phi_{i,j}|^8 < \infty$ for some $r \geq 0$, for $i = 1, \dots, N$.

Assumption 4 $\Sigma_u \equiv E(u_{t_h} u_{t_h}') = (\sigma_{u,ij})_{i,j=1,\dots,N}$, with $u_{t_h} = (u_{1,t_h}, \dots, u_{N,t_h})'$, for all t_h, i, j and is such that $\lambda_{\min}(\Sigma_u) > c_1$ and $\lambda_{\max}(\Sigma_u) < c_2$ for some positive constants c_1 and c_2 .

Assumption 5 As $N, T_H \rightarrow \infty$ such that $\log N/T_H \rightarrow 0$,

$$(a) \max_{i,j \leq N} \left| \frac{1}{T_H} \sum_{t_h=1}^{T_H} u_{i,t_h} u_{j,t_h} - \sigma_{u,ij} \right| = O_p \left(\sqrt{\frac{\log N}{T_H}} \right).$$

$$(b) \max_{i \leq N} \left\| \frac{1}{T_H} \sum_{t_h=1}^{T_H} f_{t_h} u_{i,t_h} \right\| = O_p \left(\sqrt{\frac{\log N}{T_H}} \right).$$

Assumptions 1 and 2 are similar to the Assumptions 6 and 7 in GP (2014), except that we need higher moments in Assumption 1. We require a large number of moments because our proof relies on repeated applications of Cauchy-Schwarz's inequality to prove

the validity of our bootstrap method under cross-sectional and serial dependence. If we further assume that the factors, factor loadings, and idiosyncratic error terms are mutually independent, then having $E\|\lambda_i\|^8 \leq M$, $E\|f_{t_h}\|^8 \leq M$, and $E|e_{i,t_h}|^{16} \leq M$ are sufficient. Assumption 2 justifies that we use wild bootstrap in the second step as the regression error term is a martingale difference sequence. This assumption can be relaxed to allow for serial correlation in the regression error term and block-based bootstrap can be applied as explained in Djogbenou et al. (2015). Furthermore, in Assumption 3, we assume that idiosyncratic error term is a stationary autoregressive (AR) process of infinite order with polynomial decaying coefficients. In the proof of Section 3 (see Section C in Online Appendix), we show that $r = 4$ is sufficient. Finally, Assumption 4 and Assumption 5 are similar to the CS and TS assumptions in Gonçalves and Perron (2020) (on the idiosyncratic error terms) and Gonçalves et al. (2024) (on the innovations of the idiosyncratic error terms). We assume that the variance-covariance matrix of the innovation terms is time-invariant and the innovation terms are weakly dependent in cross-sectional dimension. Under these additional assumptions, we show the validity of the AR-sieve +CSD bootstrap method in the following theorem.

Theorem 3.1 *Suppose that autoregressive sieve with CSD (AR-sieve + CSD) bootstrap and wild bootstrap are used to generate $\{e_{i,t_h}^*\}$ and $\{\varepsilon_t^*\}$, respectively with $E^*|\eta_{i,t_h}|^4 < C$ for all (i, t_h) and $E^*|\nu_t|^{4q} < C$ for all t , for some $q > 1$. If Assumptions A.1 - A.6 in Section A in Online Appendix and Assumptions 1 - 5 hold,*

$$\sup_{x \in \mathbb{R}^{r+p}} |P^*(\sqrt{T}(\Phi_0^* \tilde{\alpha}^* - \tilde{\alpha}) \leq x) - P(\sqrt{T}(\tilde{\alpha} - \alpha) \leq x)| \xrightarrow{p} 0,$$

where $\Phi_0^* = \text{diag}(1, H_0^*, I_p)$ with $H_0^* = \text{plim } H^*$ and $H^* = \tilde{V}^{*-1} \frac{\tilde{f}'^* \tilde{f}}{T_H} \frac{\tilde{\Lambda}' \tilde{\Lambda}}{N}$, which is a bootstrap analogue of rotation matrix, H .

4 Monte Carlo Simulation

In this section, we confirm the presence of bias in the factor-MIDAS regression models, and show the finite sample performance of both inference methods we propose. The data generating process (DGP) is similar to GP (2014) and Aastveit, Foroni, and Ravazzolo (2017). We consider the factor-MIDAS regression model with a single factor model as follows.

$$y_t = \beta_0 + \beta_1 \sum_{k=1}^K w_k(\theta) f_{t-k/m} + \varepsilon_t, \quad (16)$$

$$X_{i,t-k/m} = \lambda_i f_{t-k/m} + e_{i,t-k/m}, \quad k = m - 1, \dots, 0. \quad (17)$$

For a weighting function, $w_k(\theta)$, for $k = 1, \dots, K$, we use the exponential Almon lag with two parameters, (2).

The factors and factor loadings are generated similarly to GP (2014). The single factor f_t is randomly drawn from a standard normal distribution independently over time. The factor loading, λ_i is randomly drawn from a uniform distribution of the interval $[0, 1]$ independently across indicators, i . We consider that the high-frequency variable is observed at most 3 times between $t - 1$ and t (equivalent to low-frequency data being quarterly and high-frequency data being monthly), which implies $m = 3$. The parameters are $\beta_0 = 0$, $\beta_1 = 2.5$, $\theta_1 = 0.007$, and $\theta_2 = -0.01$. We choose the weighting parameters similar to Aastveit et al. (2017) to induce fast-decaying weights.

Table 1 shows six different scenarios to generate the idiosyncratic error terms and MIDAS regression error terms. We consider the error term in the regression model to be either homoskedastic or heteroskedastic. In DGP 1, we consider homoskedastic error term and in the rest of the DGPs, the error terms are conditionally heteroskedastic. When they are homoskedastic, the errors are drawn independently and identically from a standard nor-

Table 1: Data generating process

DGP	ε_t	e_{i,t_h}
1	$N(0, 1)$	$N(0, 1)$
2	$\varepsilon_t = \sqrt{h_t}v_t$	$N(0, 1)$
3	$\varepsilon_t = \sqrt{h_t}v_t$	$N(0, \sigma_i^2)$
4	$\varepsilon_t = \sqrt{h_t}v_t$	AR + $N(0, \sigma_i^2)$
5	$\varepsilon_t = \sqrt{h_t}v_t$	CS + $N(0, 1)$
6	$\varepsilon_t = \sqrt{h_t}v_t$	CS + AR

where $h_t = 0.1 + 0.3\varepsilon_{t-1}^2 + 0.6h_{t-1}$ and $v_t \sim \text{i.i.d.}N(0, 1)$ for $t = 1, \dots, T$ and $t_h = 1, \dots, T_H$.

mal distribution. To allow for heteroskedasticity, we assume that the error terms follow a GARCH model, which implies that they are conditionally heteroskedastic but unconditionally homoskedastic. Particularly, we use the same process as in Aastveit et al. (2017): $\varepsilon_t = \sqrt{h_t}v_t$, where $h_t = 0.1 + 0.3\varepsilon_{t-1}^2 + 0.6h_{t-1}$ and $v_t \sim \text{i.i.d.}N(0, 1)$.

For the idiosyncratic term in the factor model, we use the same data-generating process in GP (2014). In DGP 1 and DGP 2, the idiosyncratic error terms are homoskedastic by randomly generating them from a standard normal distribution. DGP 3 induces heteroskedasticity in the idiosyncratic term, where the variance for each indicator is drawn from $U[0.5, 1.5]$. DGP 4 introduces the serial correlation by generating the idiosyncratic term from an autoregressive model of order one such that $e_{i,t_h} = \rho_i e_{i,t_h-1} + u_{i,t_h}$, where $u_{i,t_h} \sim \text{i.i.d.}N(0, 1)$. For simplicity, we let $\rho_i = \rho$ for all $i = 1, \dots, N$, and $\rho = 0.5$. The idiosyncratic terms are re-scaled by $(1 - \rho^2)^{1/2}$ so that the variance of the idiosyncratic error terms is 1. DGP 5 allows for cross-sectional dependence in the homoskedastic idiosyncratic terms as in GP (2014) and Bai and Ng (2006). Precisely, we let the correlation between e_{i,t_h} and e_{j,t_h} be $0.5^{|i-j|}$ for $|i - j| \leq 5$ and 0 for otherwise. In DGP 6, the idiosyncratic

error terms have both serial and cross-sectional dependence. The idiosyncratic error terms follow the autoregressive process of order 1 with the innovation term being cross-sectionally correlated. The idiosyncratic terms in DGP 5 and 6 are also re-scaled to have the variance 1, the same as in other designs.

To focus on the bias, which arises by the fact that the factors are estimated, we do not estimate the number of the factors in the estimation process. Instead, we assume that we know that there is a single factor. We report the size of the bias in a slope coefficient for the single factor, β_1 . Mainly, we report two sets of results: based on asymptotic theory and based on the bootstrap method. The bias based on asymptotic theory is reported when we use the true factor, the estimated factor, and the plug-in bias estimator. We also impose that we know $Cov(e_{i,t_h}, e_{i,t_h-k}) = 0$ for $k > 1$, and therefore we only compute the plug-in bias estimator up to the first degree covariance term. The other set of results includes the bias based on two different bootstrap methods: wild bootstrap (WB) and AR-sieve+CSD bootstrap. For AR-sieve+CSD bootstrap, we choose a lag order for each series by AIC. Note that the wild bootstrap is only valid when the idiosyncratic error terms do not have serial and cross-sectional dependence, DGP 1 - 3. For the rest of the designs, the wild bootstrap is not valid. Therefore, under more general settings (DGP 4 - 6), we can quantify the cost of not accounting for either time-series or cross-sectional dependence or both in the idiosyncratic error term by comparing two bootstrap methods.

To compute the size of bias, we use the approach described in GP (2014). The bias in the original sample is calculated as the average of $H\tilde{\beta}_1 - \beta_1$. This guarantees each estimator in the replication to be consistent for β_1 . In the bootstrap world, similarly, we compute the bias of the bootstrap estimator as the average of $HH^*\tilde{\beta}_1^* - H\tilde{\beta}_1$. We also report the 95% coverage rate for the associated estimators: estimated factors, plug-in bias and two

bootstrap methods. The coverage rates associated with the bootstrap methods are reported by using the bootstrap equal-tailed percentile- t method.

All our simulation results are based on 5000 replications and 399 bootstraps. We consider $T = 50, 100, 200$ and $N = 50, 100, 200$. Since the high frequency variable is observed $m = 3$ times more, the time-series dimensions in the factor model as 150, 300, and 600, respectively. We choose $K = 11$, which implies that a low-frequency variable can be explained by 11 lagged monthly factors.

Since the results of DGP 1 - 3 are very similar, we leave the results of DGP 1 - 2 in the Online Appendix. The results of DGP 3 and 4 are presented in Table 2. In both scenarios, the MIDAS regression error terms are now heteroskedastic for both DGPs. The idiosyncratic error terms are heteroskedastic in DGP 3. We find that there exists a bias when we use the estimated factor and the plug-in estimator overestimates the magnitude of the bias, especially in small samples. Both bootstrap methods outperform the plug-in estimator in terms of replicating the bias size and correcting the distortion. In DGP 4, the idiosyncratic error terms exhibit not only heteroskedasticity but also display serially dependence. In contrast to DGP 3, the bias size increases as we introduce serial dependence in the error term of the factor model, and it is about twice as large as that in DGP 3. This is consistent with the asymptotic bias result in Theorem 2.1, where time-series dependence contributes to the bias. The plug-in bias is no longer overestimating the bias size.¹³

Comparing the two bootstrap methods, it is evident that AR-sieve+CSD bootstrap method performs better than the wild bootstrap method in DGP 4 - 6. Note that the wild bootstrap is no longer valid under serial dependence. In fact, for some sample sizes, the

¹³It is important to note that since the bias depends on the serial dependence, the persistence in the idiosyncratic error term may also have an impact. We have observed that with an increase in persistence, the bias also increases (documented in Table 1 in Section D in Online Appendix).

Table 2: DGP 3 & DGP 4 - Bias and coverage rate of 95% CIs for β

		$N = 50$			$N = 100$			$N = 200$		
		$T = 50$	100	200	50	100	200	50	100	200
		$T_H = 150$	300	600	150	300	600	150	300	600
		bias								
DGP 3: hetero & hetero	True Factor	0.00	-0.01	0.00	-0.01	0.00	0.00	0.01	0.00	0.00
	Estimated Factor	-0.37	-0.34	-0.32	-0.22	-0.19	-0.17	-0.12	-0.11	-0.10
	Plug-in	-0.41	-0.36	-0.35	-0.22	-0.20	-0.19	-0.11	-0.11	-0.10
	WB	-0.27	-0.26	-0.26	-0.17	-0.16	-0.15	-0.11	-0.10	-0.09
	AR-sieve+CSD	-0.26	-0.26	-0.25	-0.17	-0.16	-0.15	-0.11	-0.10	-0.09
		95% coverage rate								
	Estimated Factor	75.0	72.6	63.9	85.0	85.5	84.4	88.5	90.3	91.0
	Plug-in	80.9	87.9	88.9	86.8	89.3	92.1	88.9	91.1	92.5
	WB	91.7	94.2	92.7	92.6	93.5	94.1	91.3	93.9	93.8
	AR-sieve+CSD	93.7	92.1	90.4	93.6	94.3	94.1	94.1	95.1	93.6
		bias								
DGP 4: hetero & AR	True Factor	0.00	0.00	0.00	-0.01	0.00	0.00	-0.01	0.00	0.00
	Estimated Factor	-0.64	-0.57	-0.54	-0.41	-0.35	-0.31	-0.28	-0.21	-0.18
	Plug-in	-0.45	-0.42	-0.41	-0.26	-0.26	-0.25	-0.14	-0.14	-0.14
	WB	-0.22	-0.22	-0.22	-0.15	-0.14	-0.14	-0.10	-0.09	-0.08
	AR-sieve+CSD	-0.38	-0.37	-0.36	-0.29	-0.26	-0.25	-0.22	-0.18	-0.16
		95% coverage rate								
	Estimated Factor	52.2	44.5	29.2	72.3	71.8	67.3	81.5	85.0	84.1
	Plug-in	72.0	77.1	77.1	81.1	86.0	87.9	85.0	90.1	91.3
	WB	82.8	79.4	68.7	89.0	88.8	86.1	89.6	92.4	91.3
	AR-sieve+CSD	88.7	87.4	81.4	91.9	91.9	91.3	93.6	94.9	93.5

In DGP 3, both error terms are heteroskedastic. In DGP 4, the idiosyncratic error term is generated as the autoregressive process of lag 1 for each variable and with heteroskedastic. For coverage rates, the results for estimated factors and plug-ins are based on asymptotic theory. The bootstrap coverage rates use the bootstrap equal-tailed percentile t method.

wild bootstrap even performs worse than the plug-in bias, when it comes to compare the size of the bias. We can also confirm that the AR-sieve+CSD bootstrap procedure outperforms the plug-in bias and wild bootstrap procedure by comparing the results of coverage rates,

particularly in small sample sizes.

Table 3: DGP 5 & DGP 6 - Bias and coverage rate of 95% CIs for β

		$N = 50$			$N = 100$			$N = 200$		
		$T = 50$	100	200	50	100	200	50	100	200
		$T_H = 150$	300	600	150	300	600	150	300	600
		bias								
DGP 5: hetero & CSD	True Factor	0.00	-0.01	0.00	-0.01	0.00	0.00	0.01	0.00	0.00
	Estimated Factor	-0.37	-0.34	-0.32	-0.22	-0.19	-0.17	-0.12	-0.11	-0.10
	Plug-in	-0.41	-0.36	-0.35	-0.22	-0.20	-0.19	-0.11	-0.11	-0.10
	WB	-0.10	-0.10	-0.10	-0.06	-0.06	-0.04	-0.04	-0.04	-0.03
	AR-sieve+CSD	-0.16	-0.16	-0.16	-0.10	-0.10	-0.10	-0.06	-0.06	-0.06
		95% coverage rate								
	Estimated Factor	75.0	72.6	63.9	85.0	85.5	84.4	88.5	90.3	91.0
	Plug-in	80.9	87.9	88.9	86.8	89.3	92.1	88.9	91.1	92.5
	WB	88.7	86.2	79.5	92.7	92.6	90.0	94.2	93.5	93.5
	AR-sieve+CSD	90.9	90.0	87.0	93.3	94.1	92.3	94.3	93.9	93.7
		bias								
DGP 6: hetero & CSD+AR	True Factor	0.00	0.00	0.00	-0.01	0.00	0.00	-0.01	0.00	0.00
	Estimated Factor	-0.64	-0.57	-0.54	-0.41	-0.35	-0.31	-0.28	-0.21	-0.18
	Plug-in	-0.45	-0.42	-0.41	-0.26	-0.26	-0.25	-0.14	-0.14	-0.14
	WB	-0.08	-0.09	-0.08	-0.06	-0.06	-0.05	-0.04	-0.03	-0.03
	AR-sieve+CSD	-0.23	-0.23	-0.24	-0.17	-0.16	-0.16	-0.12	-0.10	-0.10
		95% coverage rate								
	Estimated Factor	52.2	44.5	29.2	72.3	71.8	67.3	81.5	85.0	84.1
	Plug-in	72.0	77.1	77.1	81.1	86.0	87.9	85.0	90.1	91.3
	WB	76.5	66.2	47.4	87.5	84.2	77.6	91.1	91.5	89.3
	AR-sieve+CSD	86.3	80.0	73.5	91.0	89.8	87.1	93.2	93.2	92.6

In DGP 5 and 6, both error terms are heteroskedastic. In DGP 5, the idiosyncratic error term contains the cross-sectional dependence. In DGP 6, we impose the dependence in both dimensions for the idiosyncratic error terms. For coverage rates, the results for estimated factors and plug-in are based on asymptotic theory. The bootstrap coverage rates use the bootstrap equal-tailed percentile t method.

Finally, we present the results of DGP 5 and 6, which are shown in Table 3. In DGP 5, the idiosyncratic error term is only cross-sectionally correlated. Comparing the size of

the bias, the AR-sieve+CSD bootstrap performs better than the wild bootstrap method but worse than the plug-in bias method. The AR-sieve+CSD bootstrap method recovers the size distortion better than the plug-in method in most of the cases. The plug-in estimation method performs better than the AR-sieve+CSD bootstrap method when $N = 50$ and $T = 200$. In DGP 6, we allow for cross-sectional dependence as well as serial dependence in the idiosyncratic error terms. The results follow a similar pattern to the findings of DGP 5. The plug-in bias method replicates the bias better than bootstrap methods. However, it does worse than AR-sieve+CSD bootstrap in terms of recovering the size distortion in the coverage rates except when $T = 200$. Furthermore, when the time series dimension is as small as 50, the plug-in bias method performs even worse than the wild bootstrap method, which is not valid in this design. Overall, the AR-sieve+CSD bootstrap works well in correcting the distortion.¹⁴

5 Empirical Application

In this section, we apply the factor-MIDAS regression model to validate the presence of bias in an empirical example. It is well documented that incorporating high-frequency indicators to forecast a quarterly variable using the MIDAS regression model improves the forecast performance (e.g., see Clements and Galvão (2008; 2009), Aastveit et al. (2017), Marcellino and Schumacher (2010), Andreou, Ghysels, and Kourtellos (2013), and Beyhum and Striaukas (2024)).

In this paper, we focus on nowcasting quarterly U.S. real GDP growth using monthly

¹⁴Similar findings can be found when the AR-sieve+CSD bootstrap is used in the context of the unrestricted MIDAS regression model. The performance of AR-sieve+CSD bootstrap dominates the plug-in bias estimation method in all DGPs. See Table 4 - 6 in Section D in Online Appendix.

macroeconomic factors from 1984 Q1 to 2022 Q4 including great moderation period. We have divided this period into two: the long period (1984 Q1 - 2022 Q4), which includes the COVID pandemic period, and the short period (1984 Q1 to 2019 Q4). Although we look into two different periods, the results are very similar; therefore, we present the results for the shorter period in the Online Appendix. Our nowcasting model is similar to the model in Beyhum and Striaukas (2024). Given the number of leading months, $l = 1, 2, 3$, we write our model as follows.

$$y_t = \beta_0 + \sum_{i=1}^{p_y} \rho_i y_{t-i} + \beta_1' \sum_{k=1-l}^{K-l} w_{(k-1)+l}(\theta) f_{t-1-(j-1)/m} + \varepsilon_t, \quad (18)$$

where y_t is quarterly U.S. GDP growth rate. We denote common factors containing timely information about monthly macroeconomic predictors by $f_{t-k/m}$. The number of leading months represents a nowcasting horizon, denoted by h . For instance, $l = 1$ indicates that we exploit information of one leading month; hence, we nowcast two months away ($h = 2$). We use the exponential Almon lag with two parameters defined in (2) for the lag polynomial function. The quarterly U.S. output is obtained from a FRED-QD dataset (for detail, see M. McCracken and Ng (2020)). As U.S. real output is available in level in the dataset, we compute the growth rate in percentage, by $\{\ln(\text{GDP})_t - \ln(\text{GDP})_{t-1}\} \times 100$. We also include the lags of the growth rate in the regression. The number of lags of the dependent variable is chosen by BIC, before we apply MIDAS regression. BIC selects one lag in the long period and three lags in the short period.

To estimate the monthly factors, we utilize the FRED-MD dataset¹⁵ (for detail, see M. W. McCracken and Ng (2016)). We consider 74 macroeconomic variables available for the entire period and exclude all financial variables. Using PCA, we extract two common factors

¹⁵We use the ‘current’ version downloaded on October 3rd, 2023.

in both periods. The information criterion proposed by Bai and Ng (2002) (particularly, IC_p) chooses eight factors in the long period and five factors in the short period. Although the information criterion chooses more than 2 factors, the two factors we extract explain more than 60% of the variability explained by all the factors chosen by the information criterion proposed by Bai and Ng (2002).

Our primary goal is to verify the existence of bias in the estimators. Instead of focusing solely on the forecasting performance of the factor-MIDAS regression model, we aim to examine the behaviour of the estimators, particularly their 90% confidence interval. We present three sets of confidence intervals, one based on asymptotic theory and the other two based on the bootstrap method. We use two different bootstrap methods for resampling the idiosyncratic error terms in the factor model: wild bootstrap and AR-sieve + CSD bootstrap, described in Section 3. We also rotate the bootstrap estimators, $\tilde{\beta}_1^*$, with the rotation matrix H^* as in GP (2014) and Gonçalves and Perron (2020).

In Table 4, we present the confidence interval for the point estimates in the long period, 1984 Q1 - 2022 Q4 for each nowcasting horizon, $h = 2, 1$, and 0. We also report the estimate associated with each parameter on the top of the three confidence intervals. We can find that there exists a bias in the estimators associated with the factors. For example, the point estimate associated with the first factor for horizon $h = 2$ is 2.54. The confidence interval of this estimate is centered around 2.54, but the bootstrap interval shifts to the right, suggesting a negative bias. The results are similar for the other horizons, $h = 1$ and 0. Although the second factor is not significant at $h = 2$, we can confirm that there exists a bias in the estimator associated with the second factor at $h = 1$ and $h = 0$. When $h = 1$, the result implies a negative bias, whereas when $h = 0$, there exists a positive bias, shifting the interval to the left. Comparing the two bootstrap methods, there is a small change in the

Table 4: Estimates in the long period (1984 Q1 - 2022 Q4)

		$h = 2$		$h = 1$		$h = 0$	
constant	Asymptotic WB AR sieve+CSD	0.90		0.83		0.99	
		0.67	1.01	0.67	0.99	0.78	1.21
		0.71	0.98	0.69	0.95	0.73	1.28
		0.71	0.98	0.69	0.94	0.75	1.26
first factor	Asymptotic WB AR sieve+CSD	2.54		3.79		1.87	
		1.64	3.44	2.97	4.61	0.31	3.44
		2.01	3.56	3.29	4.72	0.91	3.93
		2.13	3.54	3.34	4.80	0.90	3.39
second factor	Asymptotic WB AR sieve+CSD	0.04		0.36		-0.95	
		-0.22	0.30	0.08	0.65	-1.47	-0.43
		-0.17	0.37	0.14	0.75	-1.62	-0.01
		-0.12	0.38	0.16	0.77	-1.63	-0.21
y_{t-1}	Asymptotic WB AR sieve+CSD	-0.30		-0.30		-0.58	
		-0.54	-0.06	-0.52	-0.09	-0.87	-0.28
		-0.49	-0.12	-0.44	-0.14	-1.25	-0.26
		-0.49	-0.12	-0.43	-0.14	-1.22	-0.25

The interval based on the asymptotic theory is obtained by adding and subtracting 1.645 times the heteroskedasticity robust standard errors. The confidence intervals based on bootstrap methods are obtained with equal-tailed bootstrap intervals with a bootstrap number 4999. WB indicates that we use wild bootstrap and AR sieve + CSD indicates that we use the bootstrap algorithm described in Section 3.

bootstrap confidence intervals of the estimators associated with the two factors. However, the difference is not huge, indicating that the serial and cross-sectional dependence in this example may be small.

6 Conclusion

In this paper, we derive the asymptotic distribution of the estimators in the factor-augmented MIDAS regression models. We find that there exists an asymptotic bias arising from the fact

that the factors are latent and must be estimated. We show that the bias depends on the serial dependence as well as the cross-sectional dependence of the idiosyncratic error term in the factor model, because MIDAS temporally aggregates the factors and their lags. We propose two inference methods that account for this bias: an analytical bias estimator based on the bias formula derived and a bootstrap method. Both inference methods are robust to serial and cross-sectional dependence.

Although our simulation results support the theoretical findings, the bootstrap method more effectively corrects the size distortion in the coverage rates, while the plug-in method outperforms the bootstrap method in estimating the size of the bias, especially in small samples. We further apply the factor-MIDAS regression model to nowcast quarterly U.S. GDP growth rate using monthly macroeconomic factors. Our empirical results indicate that there exists a bias in the estimates associated with the estimated factors.

Our results can be extended to the context of forecasting, such as to construct forecast intervals, similar to Gonçalves, Perron, and Djogbenou (2017), where they construct it in the context of the factor-augmented regression models without mixed frequency datasets. By letting $\hat{y}_{T+1} = g(\tilde{F}_T, \tilde{\alpha})$ be the forecast of y_{T+1} based on information up to time T , we can decompose the forecast error as

$$\hat{y}_{T+1} - y_{T+1} = -\varepsilon_{T+1} + \frac{1}{\sqrt{T}} \frac{\partial g(\tilde{F}_t, \alpha)}{\partial \alpha'} \sqrt{T}(\tilde{\alpha} - \alpha) + \frac{1}{\sqrt{N}} \beta' H^{-1} \sqrt{N}(\tilde{F}_t(\theta) - H F_t(\theta)) + o_p(1).$$

This underscores the importance of the asymptotic distribution of the estimators derived in this paper in constructing to construct the forecast interval. We leave this for future research.

An interesting extension involves the use of machine-learning techniques. Machine learning techniques are popularly used to handle high-dimensional data. Along the same lines,

Babii, Ghysels, and Striaukas (2022) propose a machine learning regression by applying the sparse-group LASSO technique for mixed-frequency data.

Acknowledgement

We would like to thank John Galbraith, Eric Ghysels, Sílvia Gonçalves, and Benoit Perron for the advice. We also want to thank participants at the Canadian Economic Association (2022), International Symposium on Econometrics Theory and Applications (2022), and Canadian Econometrics Study Group (2023).

Computations were made on the supercomputer Beluga from McGill University, managed by Calcul Québec and the Digital Research Alliance of Canada.

References

- Aastveit, K. A., Foroni, C., & Ravazzolo, F. (2017). Density forecasts with midas models. *Journal of Applied Econometrics*, *32*(4), 783-801.
- Andreou, E., Gagliardini, P., Ghysels, E., & Rubin, M. (2019). Inference in group factor models with an application to mixed-frequency data. *Econometrica*, *87*(4), 1267–1305.
- Andreou, E., Ghysels, E., & Kourtellos, A. (2010). Regression models with mixed sampling frequencies. *Journal of Econometrics*, *158*(2), 246-261.
- Andreou, E., Ghysels, E., & Kourtellos, A. (2013). Should macroeconomic forecasters use daily financial data and how? *Journal of Business & Economic Statistics*, *31*(2), 240–251.

- Babii, A., Ghysels, E., & Striaukas, J. (2022). Machine learning time series regressions with an application to nowcasting. *Journal of Business & Economic Statistics*, *40*(3), 1094–1106.
- Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, *71*(1), 135–171.
- Bai, J., & Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, *70*(1), 191–221.
- Bai, J., & Ng, S. (2006). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica*, *74*(4), 1133–1150.
- Bai, J., & Ng, S. (2013). Principal components estimation and identification of static factors. *Journal of Econometrics*, *176*(1), 18–29.
- Beyhum, J., & Striaukas, J. (2024). Testing for sparse idiosyncratic components in factor-augmented regression models. *Journal of Econometrics*, *244*(1), 105845.
- Bi, D., Shang, H. L., Yang, Y., & Zhu, H. (2021). *Ar-sieve bootstrap for high-dimensional time series*. Retrieved from <https://arxiv.org/abs/2112.00414>
- Bühlmann, P. (1997). Sieve bootstrap for time series. *Bernoulli*, 123–148.
- Clements, M. P., & Galvão, A. B. (2008). Macroeconomic forecasting with mixed-frequency data: Forecasting output growth in the united states. *Journal of Business & Economic Statistics*, *26*(4), 546–554.
- Clements, M. P., & Galvão, A. B. (2009). Forecasting us output growth using leading indicators: An appraisal using midas models. *Journal of Applied Econometrics*, *24*(7), 1187–1206.
- Djogbenou, A., Gonçalves, S., & Perron, B. (2015). Bootstrap inference in regressions with estimated factors and serial correlation. *Journal of Time Series Analysis*, *36*(3),

481–502.

- Ferrara, L., & Marsilli, C. (2019). Nowcasting global economic growth: A factor-augmented mixed-frequency approach. *The World Economy*, 42(3), 846–875.
- Foroni, C., Marcellino, M., & Schumacher, C. (2015). Unrestricted mixed data sampling (midas): Midas regressions with unrestricted lag polynomials. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 178(1), 57–82.
- Ghysels, E., Santa-Clara, P., & Valkanov, R. (2004, 06). The midas touch: mixed data sampling regression.
- Ghysels, E., Santa-Clara, P., & Valkanov, R. (2005). There is a risk-return trade-off after all. *Journal of Financial Economics*, 76(3), 509–548.
- Ghysels, E., Santa-Clara, P., & Valkanov, R. (2006). Predicting volatility: getting the most out of return data sampled at different frequencies. *Journal of Econometrics*, 131(1-2), 59–95.
- Ghysels, E., Sinko, A., & Valkanov, R. (2007, 02). Midas regressions: Further results and new directions. *Econometric Reviews*, 26, 53–90.
- Ghysels, E., Valkanov, R. I., & Serrano, A. R. (2009). Multi-period forecasts of volatility: Direct, iterated, and mixed-data approaches. In *Efa 2009 bergen meetings paper*.
- Gonçalves, S., & Perron, B. (2014). Bootstrapping factor-augmented regression models. *Journal of Econometrics*, 182(1), 156–173.
- Gonçalves, S., & Perron, B. (2020). Bootstrapping factor models with cross sectional dependence. *Journal of Econometrics*, 218(2), 476–495.
- Gonçalves, S., Perron, B., & Djogbenou, A. (2017). Bootstrap prediction intervals for factor models. *Journal of Business & Economic Statistics*, 35(1), 53–69.
- Gonçalves, S., Koh, J., & Perron, B. (2024, 11). Bootstrap inference for group factor models.

Journal of Financial Econometrics, nbae020.

- Kim, H. H., & Swanson, N. R. (2018). Methods for backcasting, nowcasting and forecasting using factor-midas: With an application to korean gdp. *Journal of Forecasting*, 37(3), 281–302.
- Kock, A. B., & Callot, L. (2015). Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics*, 186(2), 325–344. (High Dimensional Problems in Econometrics)
- Krampe, J., Kreiss, J.-P., & Paparoditis, E. (2021). Bootstrap based inference for sparse high-dimensional time series models. *Bernoulli*, 27(3), 1441 – 1466.
- Kreiss, J.-P., Paparoditis, E., & Politis, D. N. (2011). On the range of validity of the autoregressive sieve bootstrap. *The Annals of Statistics*, 39(4), 2103–2130.
- Ludvigson, S. C., & Ng, S. (2009). *A factor analysis of bond risk premia* (Tech. Rep.). National Bureau of Economic Research.
- Marcellino, M., & Schumacher, C. (2010). Factor midas for nowcasting and forecasting with ragged-edge data: A model comparison for german gdp. *Oxford Bulletin of Economics and Statistics*, 72(4), 518–550.
- McCracken, M., & Ng, S. (2020). *Fred-qd: A quarterly database for macroeconomic research* (Tech. Rep.). National Bureau of Economic Research.
- McCracken, M. W., & Ng, S. (2016). Fred-md: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4), 574–589.
- Meyer, M., & Kreiss, J.-P. (2015). On the vector autoregressive sieve bootstrap. *Journal of Time Series Analysis*, 36(3), 377–397.
- Monteforte, L., & Moretti, G. (2013). Real-time forecasts of inflation: The role of financial variables. *Journal of Forecasting*, 32(1), 51–61.