A penalized two-pass regression to predict stock returns with time-varying risk premia

Gaetan Bakalli[a] and Stéphane Guerrier[b] and Olivier Scaillet[c,*]

November 2022

Abstract

We develop a penalized two-pass regression with time-varying factor loadings. The penalization in the first pass enforces sparsity for the time-variation drivers while also maintaining compatibility with the no-arbitrage restrictions by regularizing appropriate groups of coefficients. The second pass delivers risk premia estimates to predict equity excess returns. Our Monte Carlo results and our empirical results on a large cross-sectional data set of US individual stocks show that penalization without grouping can yield to nearly all estimated time-varying models violating the no-arbitrage restrictions. Moreover, our results demonstrate that the proposed method reduces the prediction errors compared to a penalized approach without appropriate grouping or a time-invariant factor model.

Keywords: two-pass regression, predictive modeling, large panel, factor model, LASSO penalization.

JEL classification: C13, C23, C51, C52, C53, C55, C58, G12, G17.

[*] Corresponding author

[a] Emlyon Business School, 23 Av. Guy de Collongue, 69130 Ecully, France. Email:bakalli@em-lyon.com

[b] Geneva School of Economics and Management and Faculty of Science, University of Geneva, 24 rue General Dufour, 1211 Geneva 4, Switzerland. Email:stephane.guerrier@unige.ch

[c] Geneva School of Economics and Management and Swiss Finance Institute, University of Geneva, 24 rue General Dufour, 1211 Geneva 4, Switzerland. Email:olivier.scaillet@unige.ch

1 Introduction

Under the arbitrage pricing theory (Ross, 1976; Chamberlain and Rothschild, 1983), we know that risk premia are drivers of expected excess returns. Hence, estimating them should be useful for prediction of future equity excess returns. The workhorse to estimate equity risk premia in a linear multi-factor setting is the two-pass cross-sectional regression method developed by Black et al. (1972) and Fama and MacBeth (1973). A series of papers address its large and finite sample properties for linear factor models with time-invariant coefficients; see, for example, Shanken (1985, 1992), Jagannathan and Wang (1998), Shanken and Zhou (2007), Kan et al. (2013), and the review paper of Jagannathan et al. (2010) (see Bryzgalova et al. (2019) for a recent

Bayesian approach). In a time-varying setting, Gagliardini et al. (2016) (henceforth referred as GOS) study how we can infer the dynamics of equity risk premia from large stock return data sets under conditional linear factor models (see also Gagliardini et al. (2020) for a review of estimation of large dimensional conditional factor models in finance). They show how to explicitly account for the no-arbitrage restrictions relating the time-varying intercept and the time-varying factor loadings when writing the underlying linear regression to be estimated. In conditional factor models, we quickly loose parsimony in terms of covariates because of the cross-products induced by the no-arbitrage restrictions. Chaieb et al. (2021) show that a direct application of the GOS methodology in an international setting is challenging because of the large number of parameters needed to model the time-variations in factor exposures and risk premia. Applying the GOS methodology off-the-shelf to an international setting results in few or even zero stocks kept for several countries. To address this issue, they suggest to rely on iteratively selecting for each stock the most important covariates driving the dynamics of the factor loadings without violating the no-arbitrage restrictions.

The aim of this paper is to tackle this issue via LASSO-type penalisation techniques (Tibshirani, 1996) to enforce sparsity for the time-variation drivers while also maintaining compatibility with the no-arbitrage restrictions. The shrinkage targets the time-invariant counterpart of the time-varying models. In a conditional factor setting, we aim at addressing the "multidimensional challenge" of Cochrane (2011), namely select characteristics which really provide independent information about average excess returns. More specifically, the penalized first-pass (time-series) regression selects and estimates the regression coefficients ensuring a model specification compatible with the no-arbitrage restrictions through the Overlap Group-LASSO (OGL) of Jacob et al. (2009) and its adaptive version of Percival (2012), the aOGL, which extends the original Group-LASSO of Yuan and Lin (2006) to groups of variables that may overlap. Indeed, if we do not introduce a quadratic term (or cross-products) in the time-varying intercept while the covariate is present in the time-varying factor loadings, we introduce *ex-ante* a model with arbitrage (see (4) below, and the discussion in Gagliardini et al., 2020). By definition, we cannot estimate a coefficient for which its covariate is absent. On the contrary, if we delete a covariate in the time-varying factor loadings and keep it in the time-varying intercept, then its corresponding coefficients could be shrunk to zero by a standard LASSO for the first-pass regression, and thus could avoid *ex-post* a model with arbitrage if the true model is sparse. In a standard Ordinary Least Squares (OLS) first-pass procedure, those time-varying intercept coefficients could be estimated close to zero if the true model does not include that covariate in the time-varying factor loadings. By introducing groups based on finance theory derived from assuming no asymptotic arbitrage opportunities in the economy, our aOGL approach can only consider models compatible *ex-ante* with the no-arbitrage restrictions by construction. The groups take explicitly into account the links between the time-varying intercept and the time-varying loadings induced by the no-arbitrage restrictions. With only models satisfying *ex-ante* the no-arbitrage restrictions, we can substantially reduce the set of possible models within our model selection procedure. We derive an upper bound, and show that the number of possible models without grouping is divided by 2^3, at least, and often by a much larger number in empirical applications. As an example, for the model specifications with four factors used in Section 5, the set of possible models satisfying *ex-ante* the no-arbitrage restrictions is 2^{97} times smaller than the set of possible models without grouping. We exemplify this reduction with a simple two-factor example in Section 3.1. It echoes the discussion in Giannone et al. (2021) that, if a prediction model with many predictors "lacks any additional structure, then there is no hope of recovering useful information about the [high-dimensional parameter] vector with limited samples" (Hastie et al., 2015, p. 290). Imposing some constraints, hopefully driven by economic reasoning (as we promote here), should help to extract relevant information in big data problems. As a consequence, the aOGL approach yields better performance in terms of covariate selection and estimated models without arbitrage (see our Monte Carlo results in Section 4 and our empirical results in Section 5). On our data for US single stocks, more than half of the stocks require dynamics in their factor loadings, while penalization without (with) grouping yields to 100% (0%) of all estimated time-varying models violating the no-arbitrage restrictions. Besides, the aOGL approach yields better in-sample and out-of-sample predictive performance on an equally-weighted portfolio (see Sections 4 and 5). On our data for US single stocks, prediction errors are located closer to zero and their scale is narrower.

LASSO type techniques have already been applied successfully to factor models in finance. Bryzgalova (2015) develops a shinkrage-based estimator that identifies the weak factors (i.e., factors that do not correlate with the assets) and ensures consistent and normality of the estimates of the risk premia. Feng et al. (2020) propose a model-selection method to evaluate the risk prices of observable factors. Freyberger et al. (2020) propose a nonparametric method to determine which firm characteristics provide incremental information for the cross section of expected excess returns. Gu et al. (2020) and Chinco et al. (2019) use penalization techniques for prediction purposes respectively at low and high-frequency. Alternatively, Fan et al. (2022) develop a nonparametric methodology for estimating conditional asset pricing models using deep neural networks, by employing time-varying conditional information on alphas and betas carried by firm-specific characteristics. Avramov et al. (2022) propose a novel Bayesian approach to study time-series and cross-sectional effects in asset returns, when the true factor model and its underlying parameters are uncertain. They use macro predictors to model time-variation in the factor loadings and investigate potential mispricing. While their prior beliefs are weighted against mispricing, their analysis shows that time-varying mispricing appears with a large probability. Chen et al. (2022) use deep neural networks to estimate a stochastic discount factor model for individual stock returns and exploit the fundamental no-arbitrage condition as criterion function, to construct the most informative test assets with an adversarial approach. Cong et al. (2022a) and Cong et al. (2022b) also use economic restrictions to enhance the performance of machine learning techniques in asset pricing and portfolio management. Fan et al. (2021) propose a new methodology that bridges the gap between sparse regressions and factor models and evaluates the gains of increasing the information set via factor augmentation to study asset returns. Finally, let us mention that there is also work on inference for large dimensional models with observable and unobservable factors with high frequency data (Fan et al., 2016; Aït-Sahalia and Xiu, 2017; Pelger and Xiong, 2019; Aït-Sahalia et al., 2020).

The outline of this paper is as follows. Section 2 describes the conditional linear factor models with sparse time-varying coefficients, and how to implement the no-arbitrage restrictions in the specification of the random coefficient panel model. Section 3 develops our penalized two-pass regression with time-varying factor loadings. The penalization in the first-pass (time-series) regressions of Section 3.1 enforces sparsity for the time-variation drivers while also maintaining compatibility *ex-ante* with the no-arbitrage restrictions through building appropriate groups of coefficients. We explain in detail in Section 3.1 why we prefer the aOGL method over the original Group-LASSO of Yuan and Lin (2006) for the first-pass regression. The second-pass (cross-sectional) regression of Section 3.2 delivers risk premia estimates to predict equity excess returns. In Section 3.2, we show asymptotic consistency of our penalised two-pass regression estimates under an adaptive estimation for the first-pass regression coefficients. Section 4 reports our simulations results. Section 5 gathers our empirical results. After describing our data on US single stocks in Section 5.1, we present our empirical results on in-sample and out-of-sample prediction performances and variable selection in Sections 5.2 and 5.3. We investigate 13 characteristics and 6 common instruments for the dynamics of factor loadings, and use the four-factor model of Carhart (1997) and the five-factor model of Fama and French (2015). Section 6 concludes. We list regularity conditions in Appendix A, the proofs of our theoretical results in Appendices B and C, and a description on how to construct groups for the numerical optimisation in Appendix D.

2 Model specification

In this section, we consider a conditional linear factor model with time-varying coefficients as in GOS (see Gagliardini et al. (2020) for a review). From their Assumptions APR.1, APR.2, and APR.3, the time-varying factor model for assets belonging to the continuum of assets $\gamma \in [0,1]$ is

$$R_t(\gamma) = a_t(\gamma) + b_t(\gamma)^\top f_t + \varepsilon_t(\gamma), \tag{1}$$

where $R_t(\gamma)$ denotes the excess return on asset γ at period $1,\dots,T$, vector $f_t \in \mathbb{R}^K$ gathers the values of the factors at date t. From Assumption APR.1 of GOS, the intercept $a_t(\gamma) \in \mathbb{R}$ and factor loadings $b_t(\gamma) \in \mathbb{R}^K$ are $\mathcal{F}_{t-1}$-measurable, where the filtration process $\mathcal{F}_{t-1}$ is the information available to all investors at time $t-1$. The error terms have mean zero $\mathbb{E}[\varepsilon_t(\gamma)|\mathcal{F}_{t-1}] = 0$ and are uncorrelated with the factors conditionally on information $\mathcal{F}_{t-1}$, $\mathrm{Cov}(\varepsilon_t(\gamma), f_{t,k}|\mathcal{F}_{t-1}) = 0$, $k = 1, ..., K$. Assumption APR.2 of GOS gathers standard measurability conditions for a stochastic process, and requires that the process $\beta_t(\gamma) = (a_t(\gamma), b_t(\gamma)^\top)^\top \in \mathbb{R}^{K+1}$ is a bounded aggregate process as defined in Al-Najjar (1995), as well as the nondegeneracy in the factor loadings across assets. Assumption APR.3 of GOS imposes an approximate factor structure in (1) such that, for any sequence $\gamma_i \in [0,1], i = 1,\dots,n$, with $\Sigma_{\varepsilon_t,t,n} \in \mathbb{R}^{n\times n}$ being the conditional variance-covariance matrix of the vector $(\varepsilon_t(\gamma_1),\dots,\varepsilon_t(\gamma_n))^\top$ knowing Z_{t-1}, there exists a set such that $n^{-1}\,\mathrm{eig}_{\text{max}}(\Sigma_{\varepsilon_t,t,n}) \xrightarrow{L^2} 0$ as $n \to \infty$, where $\mathrm{eig}_{\text{max}}(\Sigma_{\varepsilon_t,t,n})$ denotes the largest eigenvalue of $\Sigma_{\varepsilon_t,t,n}$, and where $\xrightarrow{L^2}$ denotes convergence in the L^2-norm. Under Assumptions APR.4 of GOS, the following asset pricing restriction holds:

$$a_t(\gamma) = b_t(\gamma)^\top \nu_t, \tag{2}$$

for all $\gamma \in [0,1]$, at any date $t = 1, 2, \ldots$ where random vector $\nu_t \in \mathbb{R}^K$ is unique and is $\mathcal{F}_{t-1}$-measurable, which can also be written as

$$\mathbb{E}\left[R_t(\gamma)|\mathcal{F}_{t-1}\right] = b_t(\gamma)^\top \lambda_t, \tag{3}$$

with $\lambda_t = \nu_t + \mathbb{E}[f_t|\mathcal{F}_{t-1}] \in \mathbb{R}^K$. Equation (3) shows the link between expected excess returns and the product of the time-varying factor loadings and risk premia. Below, we rely on that link to predict excess returns. Assumption APR.4 of GOS excludes asymptotic arbitrage opportunity, such that there is no portfolio sequence with zero cost and positive payoff. The conditioning information $\mathcal{F}_{t-1}$ contains $\underline{Z}_{t-1}$ and $\underline{Z}_{t-1}(\gamma)$, where $Z_{t-1} \in \mathbb{R}^p$ is a vector of lagged instruments common to all stocks, $Z_{t-1}(\gamma) \in \mathbb{R}^q$, for $\gamma \in [0,1]$, is a vector of lagged characteristics specific to stock γ, and $\underline{Z}_t = \{Z_t, Z_{t-1}, ...\}$ denotes the set of past realizations. Vector Z_{t-1} may include past observations of the factors and some additional variables such as macroeconomic variables. Vector $Z_{t-1}(\gamma)$ may include past observations of firm characteristics and stock returns. We define the dynamics of the factor loadings $b_t(\gamma)$ as a sparse linear function of Z_{t-1} (Shanken, 1990; Ferson and Harvey, 1991) and $Z_{t-1}(\gamma)$ (Avramov and Chordia, 2006).

ASSUMPTION A.1: *(Sparse time-varying factor loadings)*
The factor loadings are such that $b_t(\gamma) = A(\gamma) + B(\gamma)Z_{t-1} + C(\gamma)Z_{t-1}(\gamma)$, where $A(\gamma) \in \mathbb{R}^K$ correspond to a time-invariant model, and $B(\gamma) \in \mathbb{R}^{K\times p}$, $C(\gamma) \in \mathbb{R}^{K\times q}$ are sparse matrices of coefficient for any $\gamma \in [0,1]$ and any t.

Moreover, we define the vector of risk premia as a sparse linear function of lagged instruments Z_{t-1} (Cochrane, 1996; Jagannathan and Wang, 1996) and specify the conditional expectation of the factor $\mathbb{E}\left[f_t|\mathcal{F}_{t-1}\right]$ given the filtration process $\mathcal{F}_{t-1}$.

ASSUMPTION A.2: *(Sparse time-varying risk premia)*
The risk premia vector is such that
(i) $\lambda_t = \Lambda_0 + \Lambda_1 Z_{t-1}$, where $\Lambda_0 \in \mathbb{R}^K$ correspond to a time-invariant model and $\Lambda_1 \in \mathbb{R}^{K\times p}$ is a sparse matrix for any t.
The conditional expectation of the factor is such that
(ii) $\mathbb{E}\left[f_t|\mathcal{F}_{t-1}\right] = F_0 + F_1 Z_{t-1}$, where $F_0 \in \mathbb{R}^K$ corresponds to a time-invariant model and $F_1 \in \mathbb{R}^{K\times p}$ is a sparse matrix for any t.

Assumptions A.1 and A.2 differ from Assumptions FS.1 and FS.2 of GOS. Indeed, we consider here the matrices $B(\gamma)$, $C(\gamma)$, Λ_1 and F_1 of coefficients as sparse, meaning that only a small fraction of the Z_{t-1} or $Z_{t-1}(\gamma)$ for $\gamma \in [0,1]$ are useful to describe the dynamics of the factor loadings, risk premia, and conditional expectation of the factors. Building on the sampling scheme from Assumptions SC.1 and SC.2 of GOS, we define the indicator variable $I_t(\gamma)$, for all $\gamma \in [0,1]$, such that $I_t(\gamma) = 1$ if the return on asset γ is observable at time t, and 0 if not. Assumption SC.1 ensures that $I_t(\gamma)$, $\varepsilon_t(\gamma)$ and variables in $\mathcal{F}_{t-1}$ are independent, while Assumption SC.2 ensures that the random variables γ_i, $i = 1, ..., n$, are i.i.d. indices, independent of $\varepsilon_t(\gamma)$,

$I_t(\gamma)$, and $\mathcal{F}_{t-1}$. From the above sampling scheme, we can now use the following notation: $I_{i,t} = I_t(\gamma_i), R_{i,t} = R_t(\gamma_i), \beta_{i,t} = \beta_t(\gamma_i), \varepsilon_{i,t} = \varepsilon_t(\gamma_i), A_i = A(\gamma_i), B_i = B(\gamma_i), C_i = C(\gamma_i)$ and $Z_{i,t-1} = Z_{t-1}(\gamma_i)$ as well as $a_{i,t} = a_t(\gamma_i)$ and $b_{i,t} = b_t(\gamma_i)$. Hence, from Assumptions A.1 and A.2, we can express (1) using the asset pricing restriction in (2) as the following Data Generating Process (DGP):

$$\begin{aligned} R_{i,t} = {} & A_i^\top (\Lambda_0 - F_0) + A_i^\top (\Lambda_1 - F_1) Z_{t-1} + Z_{t-1}^\top B_i^\top (\Lambda_0 - F_0) \\ & + Z_{t-1}^\top B_i^\top (\Lambda_1 - F_1) Z_{t-1} + Z_{i,t-1}^\top C_i^\top (\Lambda_0 - F_0) \\ & + Z_{i,t-1}^\top C_i^\top (\Lambda_1 - F_1) Z_{t-1} + A_i^\top f_t + Z_{t-1}^\top B_i^\top f_t + Z_{i,t-1}^\top C_i^\top f_t + \varepsilon_{i,t}. \end{aligned} \tag{4}$$

We see that the first term $A_i^\top (\Lambda_0 - F_0)$ corresponds to the time-invariant part in the time-varying intercept $a_{i,t}$, while the term $A_i^\top f_t$ corresponds to the time-invariant part of the time-varying factor loadings $b_{i,t}$. To separate the time-invariant part from the time-varying part, we make the following assumption on the model specification.

ASSUMPTION A.3: *(Non sparse time-invariant contribution)*
We define the time-invariant contribution as $A_i^\top (\Lambda_0 - F_0) + A_i^\top f_t$. We require that the vectors $A_i \in \mathbb{R}^K, \Lambda_0 \in \mathbb{R}^K$, and $F_0 \in \mathbb{R}^K$ have a full vector specification, i.e., do not contain null-elements.

Assumption A.3 ensures that the time-invariant part of a factor loading is always included in the model specification, so that we can distinguish a factor with a time-invariant loading from a factor with a time-varying loading for asset i. This assumption is key to analyze which instrument Z_{t-1} and characteristic $Z_{i,t-1}$, if needed, drive the dynamics of the factor loadings $b_{i,t}$ for asset i, and impact on the prediction $\mathbb{E}[R_{i,t}|\mathcal{F}_{t-1}]$ via (3). Since implementing a penalized two-pass regression given on (4) is difficult (due to the quadratic form in lagged instruments Z_{t-1} and $Z_{i,t-1}$), we redefine the regressors and coefficients, as a generic panel model. Beforehand, let us define the vector of lagged instruments including the intercept as $\tilde{Z}_{t-1} = (1, Z_{t-1}^\top)^\top \in \mathbb{R}^{\tilde{p}}$, where $\tilde{p} = p + 1$, and the new matrices $\breve{B}_i = [A_i | B_i] \in \mathbb{R}^{K \times \tilde{p}}$ and $\Lambda - F = [(\Lambda_0 - F_0)|(\Lambda_1 - F_1)] \in \mathbb{R}^{K \times \tilde{p}}$ that stack respectively column-wise the elements of A_i, B_i, and $(\Lambda_0 - F_0), (\Lambda_1 - F_1)$. The linear transformed regressors are

$$x_{2,i,t} = \left(x_{21,i,t}^\top, x_{22,i,t}^\top\right)^\top = \left(f_t^\top \otimes \tilde{Z}_{t-1}^\top, f_t^\top \otimes Z_{i,t-1}^\top\right)^\top \in \mathbb{R}^{d_2},$$

where $d_2 = d_{21} + d_{22} = K\tilde{p} + Kq$, and

$$x_{1,i,t} = \left(x_{11,i,t}^\top, x_{12,i,t}^\top\right)^\top = \left(\text{vech}\left[X_t\right]^\top, \tilde{Z}_{t-1}^\top \otimes Z_{i,t-1}^\top\right)^\top \in \mathbb{R}^{d_1},$$

where $d_1 = d_{11} + d_{12} = (\tilde{p}+1)\tilde{p}/2 + \tilde{p}q$ and the symmetric matrix $X_t = (X_{t,k,l})_{k,l} \in \mathbb{R}^{\tilde{p} \times \tilde{p}}$ is such that $X_{t,k,l} = \tilde{Z}_{t-1,k}^2$, if $k = l$, and $X_{t,k,l} = 2\tilde{Z}_{t-1,k}\tilde{Z}_{t-1,l}$, otherwise, for $k, l = 1, \ldots, \tilde{p}$, where $\tilde{Z}_{t,k}$ denotes the k-th component of the vector $\tilde{Z}_t$. The vector-half operator vech $[\cdot]$ stacks the elements of the lower triangular part of a $\tilde{p} \times \tilde{p}$ matrix as a $\tilde{p}\,(\tilde{p}+1)/2$ vector. The first element of vech (X_t) is related to the time-invariant coefficients $A_i^\top (\Lambda_0 - F_0)$, whereas the elements $2, \ldots, \tilde{p}$ are related to $A_i^\top\left(\Lambda_1 - F_1\right)Z_{t-1} + Z_{t-1}^\top B_i^\top\left(\Lambda_0 - F_0\right)$. Through the above redefinitions of the regressor, we can write (4) as

$$R_{i,t} = \beta_i^\top x_{i,t} + \varepsilon_{i,t}, \tag{5}$$

where $x_{i,t} = (x_{1,i,t}^\top, x_{2,i,t}^\top)^\top$ is of dimension $d = d_1 + d_2$ and $\beta_i = (\beta_{1,i}^\top, \beta_{2,i}^\top)^\top$ is defined as

$$\begin{aligned}
\beta_{1,i} &= \left(\beta_{11,i}^\top, \beta_{12,i}^\top\right)^\top \in \mathbb{R}^{d_1},\\
\beta_{11,i} &= N_{\tilde{p}}\left[\left(\Lambda - F\right)^\top \otimes I_{\tilde{p}}\right]\operatorname{vec}[\breve{B}_i^\top] \in \mathbb{R}^{d_{11}},\\
\beta_{12,i} &= W_{\tilde{p},q}\left[\left(\Lambda - F\right)^\top \otimes I_q\right]\operatorname{vec}[C_i^\top] \in \mathbb{R}^{d_{12}},\\
N_{\tilde{p}} &= \frac{1}{2}D_{\tilde{p}}^{+}(W_{\tilde{p}} + I_{\tilde{p}^2}) \in \mathbb{R}^{[(\tilde{p}+1)\tilde{p}/2+\tilde{p}q]\times\tilde{p}^2},\\
\beta_{2,i} &= \left(\beta_{21,i}^\top, \beta_{22,i}^\top\right)^\top = \left(\operatorname{vec}[\breve{B}_i^\top]^\top, \operatorname{vec}[C_i^\top]^\top\right)^\top \in \mathbb{R}^{d_2},
\end{aligned}$$

and where $W_{\tilde{p},q}$ is the commutation matrix such that $\operatorname{vec}[M^\top] = W_{\tilde{p},q}\operatorname{vec}[M]$. Moreover, $D_{\tilde{p}}^{+}$ denotes the $((\tilde{p}+1)\tilde{p}/2+\tilde{p}q)\times\tilde{p}^2$ Moore-Penrose inverse of the duplication matrix $D_{\tilde{p}}$ such that $\operatorname{vech}[M] = D_{\tilde{p}}^{+}\operatorname{vec}[M]$, for any matrix $\tilde{p}\times\tilde{p}$ matrix M. The following section describes the selection and estimation part of the model.

3 Estimation and selection

This section implements the two-pass regression of Black et al. (1972) and Fama and MacBeth (1973), while selecting the contributing variables in the time-varying factor loadings. The penalized first-pass (time-series) regression selects and estimates the non-zero coefficients β_i for $i = 1, \dots, n$, ensuring a model specification compatible *ex-ante* with the no-arbitrage restrictions through the aOGL approach of Percival (2012). The second-pass regression relies on the Weighted Least-Square (WLS) estimator of GOS to estimate the vector ν, and takes the adaptive LASSO (aLASSO) estimator of Zou (2006) to select and estimate the matrix F of coefficients of the conditional expectation of the factors.

3.1 First-pass regression

The goal of the penalized first-pass regression is to select and estimate the factor loadings for each asset $i = 1, \dots, n$, while keeping their respective time-invariant contribution fully specified as described in Assumption A.3. Moreover, it aims at selecting variables ensuring a proper model specification consistent *ex-ante* with the no-arbitrage restrictions for each stock. A possible solution to ensure that these restrictions are satisfied while allowing to select variables in the first-pass regression is to consider a LASSO-type estimator based on appropriate predefined sets of indices corresponding to groups of variables. We define $\mathcal{G} \subset \mathcal{P}(\{1, \dots, d\})$ as the set of indices corresponding to all possible (potentially overlapping) groups in line with the no-arbitrage restrictions, where $\mathcal{P}(\{1,\ldots,d\})$ denotes the power set of $\{1,\ldots,d\}$. Moreover, we let $g \in \mathcal{G}$ denote a possible group and we require that the indices associated to all covariates belong to at least one group. Under the framework discussed in the previous sections, we define below the restrictions on $\mathcal{G}$ such that a model selection procedure based on $\mathcal{G}$ satisfies *ex-ante* the no-arbitrage restrictions by construction.

RESTRICTION R.1: *The time-invariant coefficients belong to a single group, where no amount of shrinkage is applied.*

RESTRICTION R.2: *Each covariate related to the non-diagonal elements of X_t belongs to a single group.*

RESTRICTION R.3: *For instrument $\tilde{Z}_{t-1,l}$, for $l = 1,\ldots,\tilde{p}$, if all its corresponding $\tilde{Z}_{t-1,l} f_{t,k}$, for $k = 1,\ldots,K$, in $x_{2,i,t}$ are not included in the estimated model, only the regressors $\tilde{Z}^2_{t-1,l}$, related to the diagonal element of X_t, in $x_{1,i,t}$ should not be included. For characteristic $Z_{i,t-1,m}$, for $m = 1,\ldots,q$, if all its corresponding $Z_{i,t-1,m} f_{t,k}$ for $k = 1,\ldots,K$, in $x_{2,i,t}$ are not included in the estimated model, only the regressors $Z_{i,t-1,m}$ in $x_{1,i,t}$ should not be included.*

RESTRICTION R.4: *For instrument $\tilde{Z}_{t-1,l}$, for $l = 1,\ldots,\tilde{p}$, if at least one of its corresponding $\tilde{Z}_{t-1,l} f_{t,k}$, for $k = 1,\ldots,K$, in $x_{2,i,t}$ are included in the estimated model, only the regressors $\tilde{Z}^2_{t-1,l}$, related to the diagonal element of X_t, in $x_{1,i,t}$ should be included. For characteristic $Z_{i,t-1,m}$, for $m = 1,\ldots,q$, if at least one of its corresponding $Z_{i,t-1,m} f_{t,k}$, for $k = 1,\ldots,K$, in $x_{2,i,t}$ are included in the estimated model, only the regressors $Z_{i,t-1,m}$ in $x_{1,i,t}$ should be included.*

These restrictions ensure that Assumption A.3 is satisfied and that a model selection procedure guarantees that the instrument $\tilde{Z}_{t-1,l}$ or characteristic $Z_{i,t-1,m}$ exist in either both $x_{1,i,t}$ and $x_{2,i,t}$, or neither. More specifically, Restriction R.1 is related to Assumption A.3, which requires the coefficients in β_i related to the time-invariant contribution to be always included in the selected model. Restriction R.2 is related to Assumption A.1 and Assumption A.2. Under the DGP in (4), and from the definition of $\text{vech}(X_t)$, we can see that the off-diagonal of X_t in $\text{vech}[X_t]$ cannot be assigned to any groups. We cannot assign $2\tilde{Z}_{t-1,s}\tilde{Z}_{t-1,l}$ to a group a priori, since its contribution can come from either the specification in Assumption A.1 or A.2. Restriction R.2 reflects this point, and imposes no specific group-structure to those covariates which are penalized individually. Restrictions R.3 and R.4 are critical in the model building. They constrain the set of possible models only to those compatible with the no-arbitrage restrictions, so that we do not introduce arbitrage *ex-ante* in the model specified in (5). We want to avoid that the no-arbitrage restriction $a_{i,t} = b_{i,t}^\top \nu_t$ is violated by construction *ex-ante* in the specification.

To satisfy the above restrictions, the Group-LASSO of Yuan and Lin (2006) constrains the set of possible models. For its implementation, we need to create a group with all scaled factors and their corresponding terms in the intercept, hence it implies that we select either all scaled factors (keep the group) or none of them (delete the group). To illlustrate this point, let us consider the following simple case with one common instrument, say inflation, and the Fama-French five-factor model (Fama and French, 2015). The Group-LASSO would force us to select either all scaled factors (product between lagged inflation and the factors), or none of them. It removes the possibility that only a subset of them is relevant; for example, only the product of inflation and the market factor matters for the dynamics of excess returns. Besides, we could think of using multiple groups, each one containing one scaled factor and its associated instrument. Jacob et al. (2009) investigate such a proposal and show that this approach is not appropriate as the Group-LASSO removes all groups if at least one of those groups is not selected.

To tackle this problem, Jacob et al. (2009) propose the OGL, or latent Group-LASSO. They introduce the latent variables $v_g \in \mathcal{V}_g = \{x \in \mathbb{R}^d | \operatorname{supp}(x) = g\}$, for $g \in \mathcal{G}$ and where $\operatorname{supp}(x)$ denotes the support of x, i.e., the set of indices $i \in \{1, \ldots, d\}$ such that $x_i \neq 0$. Moreover, we define $v_g = (v_{g_1}^\top, \ldots, v_{g_J}^\top)^\top \in \mathbb{R}^d$, $\mathcal{V}(\beta) = \{v_g : g \in \mathcal{G}\}$, s.t. $\beta = \sum_{g \in \mathcal{G}} v_g$, and $J = |\mathcal{G}|$, $|\cdot|$ denotes the cardinality of a set and g_j, $j = 1, \ldots, J$, denotes the j-th element of $\mathcal{G}$. Hence, the OGL estimator is the solution of the following optimization problem:

$$\hat{\beta}_i = \underset{\beta_i \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{T_i} \sum_t \left(I_{i,t} R_{i,t} - \beta_i^\top I_{i,t} x_{i,t} \right)^2 + 2\delta \|\beta_i\|_{2,1,\mathcal{G}}, \tag{6}$$

with the penalty term $\|\beta_i\|_{2,1,\mathcal{G}}$ defined as

$$\|\beta_i\|_{2,1,\mathcal{G}} = \min_{\mathcal{V}(\beta)} \sum_{g \in \mathcal{G}} \|v_g\|, \tag{7}$$

where $\|\cdot\|$ denotes the l_2-norm. In this work, we consider the adaptive version of OGL (aOGL) studied by Percival (2012), for which the estimator is described as follow:

$$\hat{\beta}_i = \underset{\beta_i \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \frac{1}{T_i} \sum_t \left(I_{i,t} R_{i,t} - \beta_i^\top I_{i,t} x_{i,t} \right)^2 + 2\delta \min_{\mathcal{V}(\beta)} \sum_{g \in \mathcal{G}} \delta_g \|v_g\| \right\}, \tag{8}$$

where $\delta_g \geq 0$ denotes the data-dependent (adaptive) weight associated to group g, and $\delta \geq 0$ corresponds to the overall amount of shrinkage. There are different strategies available in the literature for the Group LASSO and OGL to get estimator consistency and support selection consistency. They are based on the irrepresentable condition (Bach, 2008; Jacob et al., 2009), adaptive shrinkage (Nardi and Rinaldo, 2008; Percival, 2012) and group sparsity (Lounici et al., 2011). We choose adaptive shrinkage since it simplifies the presentation and derivation of our asymptotic results in a random design setting. Since our goal is to shrink toward the model that includes only the time-invariant contribution of the covariates, the weight associated with the first element of δ_g is equal to zero. The penalty term in (7) leads to a solution which is a union of the groups due to the latent variables v_g. One strategy to solve the minimization problem given in (6) and (8) is the duplication of covariates put forward in Jacob et al. (2009), that we adapt to our setting. In line with Restrictions R.1 to R.4, we consider 4 different group types. The first group includes the time-invariant intercept and time-invariant factors, and is not penalised. The second set of groups contains the covariates related to Restriction R.2, which are penalized individually. The next two sets of groups consider Restrictions R.3 and R.4. They respectively group the terms in Z_{t-1}^2 and $Z_{i,t-1}$ from $x_{1,i,t}$ with their corresponding scaled factors in $x_{2,i,t}$. The columns of the initial vector with the elements indexed by the group g, which need to be duplicated, create a new vector of duplicated regressors. Then, we can solve the optimization problem in (8) considering the duplicated regressors (instead of the initial ones), using the existing standard algorithm for the Group-LASSO. Appendix D describes in detail how to construct those groups complying with the no-arbitrage restrictions *ex-ante*, and yielding the full vector of duplicated regressors used in the numerical optimisation.

Let us now compare the number of possible models under aOGL and aLASSO methods. For the aOGL approach, we can associate a model to every subset of $\mathcal{G}$. Indeed, consider $\mathcal{W} \subseteq \mathcal{G}$, then this subset is associated to the set $S_{\mathcal{W}} = \bigcup_{l=1}^{|\mathcal{W}|} \mathcal{W}_l$ of indices. It allows us to enumerate the number 2^{J-1} of possible models under appropriate grouping. That number is typically much lower in empirical applications than the number 2^{d-n_1} of possible models with a LASSO penalization, where n_1 is the number of covariates associated to the time-invariant contribution group. We get the ratio $2^{J-1}/2^{d-n_1} = 2^{-(pq+p+q)}$, and we can see that, for large p and q, the aLASSO method examines many more possibilities. Besides, from Assumption A.1, we have $\min(p,q) \geq 1$, and deduce the upper bound:

$$\frac{2^{J-1}}{2^{d-n_1}} \leq \frac{1}{8}. \tag{9}$$

To further illustrate the grouping structure and the importance of Restrictions R.1 to R.4, let us consider the following simple two-factor model with a single common instrument and a single characteristic. Here, we have $K = 2$, $\tilde{p} = 2$, and $q = 1$, with $\tilde{Z}_{t-1} = (1, Z_{t-1})^\top \in \mathbb{R}^2$, so that the regressors $x_{i,t} = (x_{1,i,t}^\top, x_{2,i,t}^\top)^\top$ become

$$\begin{aligned} x_{1,i,t} &= (x_{1,i,t,1}, x_{1,i,t,2}, x_{1,i,t,3}, x_{1,i,t,4}, x_{1,i,t,5})^\top \\ &= (1, 2Z_{t-1}, Z_{t-1}^2, Z_{i,t-1}, Z_{t-1}Z_{i,t-1})^\top \in \mathbb{R}^5, \end{aligned}$$

and

$$\begin{aligned} x_{2,i,t} &= (x_{2,i,t,1}, x_{2,i,t,2}, x_{2,i,t,3}, x_{2,i,t,4}, x_{2,i,t,5}, x_{2,i,t,6})^\top \\ &= (f_{t,1}, Z_{t-1}f_{t,1}, f_{t,2}, Z_{t-1}f_{t,2}, Z_{i,t-1}f_{t,1}, Z_{i,t-1}f_{t,2})^\top \in \mathbb{R}^6, \end{aligned}$$

with their respective coefficients $\beta_{1,i} = (\beta_{1,i,1}, \beta_{1,i,2}, \beta_{1,i,3}, \beta_{1,i,4}, \beta_{1,i,5})^\top$ and $\beta_{2,i} = (\beta_{2,i,1}, \beta_{2,i,2}, \beta_{2,i,3}, \beta_{2,i,4}, \beta_{2,i,5}, \beta_{2,i,6})^\top$. From the definition of grouping structure in Apprendix D, we construct the set of six groups made of the covariates: $(x_{1,i,t,1}, x_{2,i,t,1}, x_{2,i,t,3})^\top$ for the time-invariant contribution, $(x_{1,i,t,2})$ for the covariate associtated to Restriction R.2, $(x_{1,i,t,3}, x_{2,i,t,2})^\top$ and $(x_{1,i,t,3}, x_{2,i,t,4})^\top$ grouping the covariates in $\tilde{Z}_{t-1}$, and finally $(x_{1,i,t,4}, x_{1,i,t,5}, x_{2,i,t,5})^\top$ and $(x_{1,i,t,4}, x_{1,i,t,5}, x_{2,i,t,6})^\top$ grouping the covariates in $\tilde{Z}_{i,t-1}$. Stacking those vectors row-wise in a single column defines the full vector of duplicated covariates for the numerical optimisation in the aOGL estimation. Besides, we can use this simple example to illustrate two possible manners to introduce *ex-ante* arbitrage through careless modeling. Removing the covariates $x_{2,i,t,2} = Z_{t-1}f_{t,1}$ and $x_{2,i,t,4} = Z_{t-1}f_{t,2}$ from the

	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	$x_{1,5}$	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	$x_{2,5}$	$x_{2,6}$
$\mathcal{M}_1$	✓	✗	✗	✗	✗	✓	✗	✓	✗	✗	✗
$\mathcal{M}_2$	✓	✓	✗	✗	✗	✓	✗	✓	✗	✗	✗
$\mathcal{M}_3$	✓	✗	✓	✗	✗	✓	✓	✓	✗	✗	✗
$\mathcal{M}_4$	✓	✗	✓	✗	✗	✓	✗	✓	✓	✗	✗
$\mathcal{M}_5$	✓	✓	✓	✗	✗	✓	✓	✓	✗	✗	✗
$\mathcal{M}_6$	✓	✓	✓	✗	✗	✓	✗	✓	✓	✗	✗
$\mathcal{M}_7$	✓	✗	✓	✗	✗	✓	✓	✓	✓	✗	✗
$\mathcal{M}_8$	✓	✓	✓	✗	✗	✓	✓	✓	✓	✗	✗
$\mathcal{M}_9$	✓	✗	✗	✓	✓	✓	✗	✓	✗	✓	✗
$\mathcal{M}_{10}$	✓	✗	✗	✓	✓	✓	✗	✓	✗	✗	✓
$\mathcal{M}_{11}$	✓	✗	✗	✓	✓	✓	✗	✓	✗	✓	✓
$\mathcal{M}_{12}$	✓	✓	✗	✓	✓	✓	✗	✓	✗	✓	✗
$\mathcal{M}_{13}$	✓	✓	✗	✓	✓	✓	✗	✓	✗	✗	✓
$\mathcal{M}_{14}$	✓	✓	✗	✓	✓	✓	✗	✓	✗	✓	✓
$\mathcal{M}_{15}$	✓	✗	✓	✓	✓	✓	✓	✓	✗	✓	✗
$\mathcal{M}_{16}$	✓	✗	✓	✓	✓	✓	✓	✓	✗	✗	✓
$\mathcal{M}_{17}$	✓	✗	✓	✓	✓	✓	✓	✓	✗	✓	✓
$\mathcal{M}_{18}$	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗
$\mathcal{M}_{19}$	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✓
$\mathcal{M}_{20}$	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓
$\mathcal{M}_{21}$	✓	✗	✓	✓	✓	✓	✗	✓	✓	✓	✗
$\mathcal{M}_{22}$	✓	✗	✓	✓	✓	✓	✗	✓	✓	✗	✓
$\mathcal{M}_{23}$	✓	✗	✓	✓	✓	✓	✗	✓	✓	✓	✓
$\mathcal{M}_{24}$	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✗
$\mathcal{M}_{25}$	✓	✓	✓	✓	✓	✓	✗	✓	✓	✗	✓
$\mathcal{M}_{26}$	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓
$\mathcal{M}_{27}$	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✗
$\mathcal{M}_{28}$	✓	✗	✓	✓	✓	✓	✓	✓	✓	✗	✓
$\mathcal{M}_{29}$	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓
$\mathcal{M}_{30}$	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗
$\mathcal{M}_{31}$	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓
$\mathcal{M}_{32}$	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Set of possible models according to Restrictions R.1-R.4 when $K = 2$, $\tilde{p} = 2$, and $q = 1$. A check denotes inclusion of a covariate in model $\mathcal{M}_j$. A cross denotes exclusion of a covariate in $\mathcal{M}_j$. For notational simplicity, we remove i and t in the column labeling such that $x_{l,i,t,k} = x_{l,k}$

full model might introduce *ex-ante* arbitrage through $x_{1,i,t,3} = Z_{t-1}^2$ since we miss its associated scaled factors in $x_{2,i,t}$. Here, the coefficient associated with $x_{1,i,t,3}$ might be shrunk to zero by the aLASSO estimator, avoiding *ex-post* a model with arbitrage. On the contrary, removing the quadratic term $x_{1,i,t,3}$, while keeping its corresponding scaled factors $x_{2,i,t,2}$ and $x_{2,i,t,4}$, introduces *ex-ante* arbitrage in the model by construction, since we cannot estimate the coefficient of $x_{1,i,t,3}$, when that covariate is absent from the model.

Table 1 lists the set $\mathcal{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_{32}\}$ of possible models that respect Restrictions R.1 to R.4 with $\mathcal{M}_1$ being the model with the time-invariant contribution only (Assumption A.3). The aOGL method gives 2^5 possible models. It is considerably smaller than the $2^8 = 256$ possible models under the aLASSO method. Here, we reach the upper bound (9) since $p = q = 1$. We can see that our regularization approach restricts the space of searched models, even in this simple time-varying setting, and hence permits a sound exploration of the possible models consistent with finance theory. Moreover, the two specifications with arbitrage described in the above lines are not in the set $\mathcal{M}$ of models induced by the grouping structure of the aOGL approach, strengthening conducive arguments for our proposed method.

Having showed the advantages of the aOGL in terms of model building, we now state the asymptotic result of the first-pass regression. Beforehand, we introduce some notations from Percival (2012). Let us define the two sets of indices $H_i = \{l \in \{1, \ldots, d\} : \beta_{i,l} \neq 0\}$, $H_i^c = \{l \in \{1, \ldots, d\} : \beta_{i,l} = 0\}$, corresponding to the sets of non-zero and zero true coefficient β_i. Moreover, we take

$$\begin{aligned} G_{H_i} &= \{g \in \mathcal{G} : g \subseteq H_i\}, \\ G_{H_i^c} &= \{g : g \subseteq H_i^c\}, \\ G_{H_{0,i}} &= \{g : |g \cap H_i| > 0; |g \cap H_i^c| > 0\}, \end{aligned}$$

the sets of groups in which the indices are respectively all non-zero, all zero and a mix of zero and non-zero in β_i. To investigate the asymptotic properties of the estimator in (8), we make the following assumptions:

ASSUMPTION A.4: $\operatorname{plim}_{T_i \to \infty} \hat{Q}_{x,i} = Q_{x,i}$*, where* $\hat{Q}_{x,i} = \frac{1}{T_i} \sum_t I_{i,t} x_{i,t} x_{i,t}^\top$ *and* $Q_{x,i} = \mathbb{E}[x_{i,t} x_{i,t}^\top | \gamma_i]$ *is positive definite.*

ASSUMPTION A.5: $\mathbb{E}[\varepsilon_{i,t} | \varepsilon_{i,\underline{t-1}}, \mathcal{F}_t] = 0$ *with* $\varepsilon_{i,\underline{t}} = \{\varepsilon_{i,t}, \varepsilon_{i,t-1}, \ldots\}$ *and there exists a positive constant* M *such that for all* n, T, $\frac{1}{M} \leq \sigma_i^2 \leq M$, $i = 1, ..., n$, *with* $\sigma_i^2 = \mathbb{E}[\varepsilon_{i,t}^2 | \gamma_i]$.

ASSUMPTION A.6: *There exists a neighborhood in* $\mathbb{R}^d$ *around* β_i *such that the decomposition of any vector* b *in the neighborhood has unique decomposition* $\{v_{i,g}^b\}$ *minimizing the norm* $\|\beta_i\|_{2,1,\mathcal{G}}$*. In particular, the decomposition* $\{v_{i,g}^b\}$*, minimizing the norm* $\|\beta_i\|_{2,1,\mathcal{G}}$ *is unique. Further, this decomposition is such that* $\{v_{i,0}^b\} = 0$*, for all* $g \in G_{H_{0,i}}$.

Assumption A.4 is a usual assumption for the standard OLS solution to be consistent (see Assumption B.1 in GOS), while Assumption A.5 allows for or a martingale difference sequence for the error terms (see Assumption A.1 in GOS). Assumption A.6 is discussed in Percival (2012) and addresses the uniqueness of decomposition of β_i. We now state the main result for the first-pass regression, which corresponds to Theorem 2 derived by Percival (2012) in the fixed design framework.

LEMMA 1: *(Asymptotic normality of $\hat{\beta}_i$)*
Under Assumptions APR.1 to APR.3, SC.1 and SC.2 of GOS, Assumptions A.4 to A.6, let β_i^{init} be an initial $\sqrt{T_i}$-consistent estimator and let $\{v_{i,g}^{init}\} = \mathcal{V}(\beta_i^{init})$ be any decomposition minimizing the norm $\|\beta_i^{init}\|_{2,1,\mathcal{G}}$. For all $i \in \{1,\dots,n\}$, let $\delta_g = {}^{1}\!/\!\|v_{i,g}^{init}\|^{\check{\gamma}}$, for $\check{\gamma} > 0$, such that $T_i^{(\check{\gamma}+1)/2}\delta \to \infty$. If $\sqrt{T_i}\delta \to 0$, then, as $T_i \to \infty$, we get the convergence in distribution:

$$\sqrt{T_i}\left(\hat{\beta}_i - \beta_i\right) \Longrightarrow V_i,$$

where the vector V_i has entries

$$V_{H_i} \sim N(0, \sigma_i^2 Q_{H_i,x,i}^{-1}),$$
$$V_{H_i^c} = 0,$$

where $Q_{H_i,x,i}$ is the submatrix of $Q_{x,i}$ with indices in H_i.

For the above result to hold, the vector β_i^{init} needs to be $\sqrt{T_i}$-consistent. More specifically, β_i^{init} is any a_{T_i}-consistent estimator where $a_{T_i} \to \infty$, and $a_{T_i}^{\check{\gamma}}\sqrt{T_i}\delta \to \infty$. Moreover, in the context of the aOGL, the decomposition $\{v_{i,g}^{\text{init}}\}$ must be unique. Lemma 4 in Percival (2012) shows $\sqrt{T_i}$-consistency of the $\{v_{i,g}^{\text{OLS}}\}$, which is a example of a potential solution for $\{v_{i,g}^{\text{init}}\}$ in the case of fixed covariates. In our framework, with the uniqueness assumption of the decomposition, we can use the ridge regression estimator as $\{v_{i,g}^{\text{init}}\}$. The distributional result of Lemma 1 is key in deriving the asymptotic properties of the second-pass regression discussed in the next section.

To control for short sample size, and potentially numerical instability on the inversion of matrix $\hat{Q}_{x,i}$, we consider the trimming device defined in GOS, such that $\mathbf{1}_i^{\chi} = \mathbf{1}\{CN(\hat{Q}_{x,i}) \leq \chi_{1,T}, \tau_{i,T} \leq \chi_{2,T}\}$, where $CN(\hat{Q}_{x,i}) = \sqrt{\text{eig}_{\max}(\hat{Q}_{x,i})/\,\text{eig}_{\min}(\hat{Q}_{x,i})}$ is the condition number of the matrix $\hat{Q}_{x,i}$, $\text{eig}_{\min}(\cdot)$ denotes the minimum eigenvalue, and $\tau_{i,T} = T/T_i$. The first trimming based on $CN(\hat{Q}_{x,i}) \leq \chi_{1,T}$ selects the assets for which the time-series regression is not badly conditioned, while the second trimming based on $\tau_{i,T} \leq \chi_{2,T}$ keeps only the assets for which samples are not too short.

3.2 Second-pass regression

The second-pass regression aims at computing the cross-sectional estimator of ν. For that purpose, we implement the WLS estimator of GOS, while accounting for the sparse model specification in the first-pass regression for all $i = 1,\dots,n$. For that purpose, we introduce the indicator vector $\mathbf{1}_{\beta_i} \in \mathbb{N}^d$, such that $\mathbf{1}_{\beta_{i,l}} = 1$ if $\beta_{i,l} \neq 0$, and 0 otherwise, for $l = 1,\dots,d$, that we decompose in the following manner: $\mathbf{1}_{\beta_i} =$

$(\mathbf{1}_{\beta_{11,i}}^\top, \mathbf{1}_{\beta_{12,i}}^\top, \mathbf{1}_{\beta_{21,i}}^\top, \mathbf{1}_{\beta_{22,i}}^\top)^\top$, where $\mathbf{1}_{\beta_{11,i}} \in \mathbb{N}^{d_{11}}$, $\mathbf{1}_{\beta_{12,i}} \in \mathbb{N}^{d_{12}}$, $\mathbf{1}_{\beta_{21,i}} \in \mathbb{N}^{d_{21}}$ and $\mathbf{1}_{\beta_{22,i}} \in \mathbb{N}^{d_{22}}$. To implement the WLS estimator for the vector ν, we need to account for the different number of regressors selected through the aOGL approach. Hence, in the same spirit as in Chaieb et al. (2021), we introduce the following selection matrices that help us transforming the $x_{i,t}$ into their sparse counterparts. The matrices $\tilde{D}_i$ and $\tilde{E}_i$ are the $d_{11} \times d_{11,i}$ and $d_{12} \times d_{12,i}$ such that columns with all zeros have been removed in $\text{diag}[\mathbf{1}_{\beta_{11,i}}]$ and $\text{diag}[\mathbf{1}_{\beta_{12,i}}]$. Similarly, the matrices $\tilde{B}_i$ and $\tilde{C}_i$ are the $d_{21,i} \times d_{21}$ and $d_{22,i} \times d_{22}$ matrices such that rows with all zeros have been removed in $\text{diag}[\mathbf{1}_{\beta_{21,i}}]$ and $\text{diag}[\mathbf{1}_{\beta_{22,i}}]$. Moreover, we define $x_{H_i,i,t}$ as the vector of regressors indexed by H_i after the selection of the first pass.

Based on the selection matrices $\tilde{D}_i, \tilde{E}_i, \tilde{B}_i$, and $\tilde{C}_i$, we rewrite the parameter restriction in (2) such that

$$\begin{aligned}
\beta_{1,i} = \Big(& \tilde{D}_i^\top N_{\tilde{p}} \left[(\Lambda - F)^\top \otimes I_{\tilde{p}} \right] \tilde{B}_i^\top \tilde{B}_i \, \text{vec} \left[\breve{B}_i^\top \right], \\
& \tilde{E}_i^\top W_{\tilde{p},q} \left[(\Lambda - F)^\top \otimes I_q \right] \tilde{C}_i^\top \tilde{C}_i \, \text{vec} \left[C_i^\top \right] \Big)^\top, \\
\beta_{3,i} = \Big(& \left[\tilde{D}_i^\top N_{\tilde{p}} \left(\breve{B}_i^\top \otimes I_{\tilde{p}} \right) \right]^\top, \left[\tilde{E}_i^\top W_{\tilde{p},q} \left(C_i^\top \otimes I_p \right) \right]^\top \Big)^\top,
\end{aligned}$$

where $N_{\tilde{p}}$ is defined in (2), yielding the asset pricing restrictions expressed in the newly defined $\beta_{1,i}$ and $\beta_{3,i}$ as $\beta_{1,i} = \beta_{3,i}\nu$, $\nu = \text{vec}[\Lambda^\top - F^\top]$. We obtain $\beta_{3,i}$ from the following identity,

$$\begin{aligned}
& \text{vec}[\beta_{3,i}^\top] = J_{a,i}\beta_{2,i}, \\
& \quad J_{a,i} = \begin{pmatrix} J_{11,i} & 0 \\ 0 & J_{22,i} \end{pmatrix}, \\
& \quad J_{11,i} = W_{d_{11,i},Kp} \left[I_{Kp} \otimes \left(\tilde{D}_i^\top N_{\tilde{p}} \right) \right] \{ I_K \otimes [(W_p \otimes I_p) (I_p \otimes \text{vec}\,[I_p])] \} \tilde{B}_i^\top, \\
& \quad J_{22,i} = W_{d_{12,i},Kp} \left[I_{Kp} \otimes \left(\tilde{E}_i^\top W_{p,q} \right) \right] \{ I_K \otimes [(W_{p,q} \otimes I_p) (I_p \otimes \text{vec}\,[I_q])] \} \tilde{C}_i^\top.
\end{aligned}$$

We can now implement the following second-pass regression WLS estimator

$$\hat{\nu} = \hat{Q}_{\beta_3}^{-1} \frac{1}{n} \sum_i \hat{\beta}_{3,i}^\top \hat{w}_i \hat{\beta}_{1,i},$$

where $\hat{\nu}$ denotes the estimator of ν, $\hat{Q}_{\beta_3} = \frac{1}{n} \sum_i \hat{\beta}_{3,i}^\top \hat{w}_i \hat{\beta}_{3,i}$, and weights are estimates of $w_i = \mathbf{1}_i^\chi \left(\text{diag}\,[v_i]\right)^{-1}$. Moreover, the v_i are the asymptotic variances of the standardized errors $\sqrt{T}(\hat{\beta}_{1,i} - \hat{\beta}_{3,i}\nu)$ in the cross-sectional regression for large T such that $v_i = \tau_i C_{\nu,1,i}^\top Q_{H_i,x,i}^{-1} S_{ii} Q_{H_i,x,i}^{-1} C_{\nu,1,i}$, where $S_{ii} = \text{plim}_{T\to\infty} \frac{1}{T} \sum_t \sigma_i^2 x_{H_i,i,t} x_{H_i,i,t}^\top$ and $C_{\nu,1,i} = (E_{1,i}^\top - (I_{d_{1,i}} \otimes \nu^\top) J_{a,i} E_{2,i}^\top)^\top$, $E_{1,i} = (I_{d_{1,i}}, 0_{d_{1,i} \times d_{2,i}})^\top$, $E_{2,i} = (0_{d_{2,i} \times d_{1,i}}, I_{d_{2,i}})^\top$. We use the estimates $\hat{v}_i = \tau_{i,T} C_{\hat{\nu}_1}^\top \hat{Q}_{H_i,x,i}^{-1} \hat{S}_{ii} \hat{Q}_{H_i,x,i}^{-1} C_{\hat{\nu}_1}$, where $\hat{S}_{ii} = \frac{1}{T_i} \sum_t I_{i,t} \hat{\varepsilon}_{i,t}^2 x_{H_i,i,t} x_{H_i,i,t}^\top$, $\hat{\varepsilon}_{i,t} = R_{i,t} - \hat{\beta}_i^\top x_{H_i,i,t}$ together with $C_{\hat{\nu},1,i} = (E_{1,i}^\top - \left(I_{d_{1,i}} \otimes \hat{\nu}_{1,i}^\top\right) J_{a,i} E_{2,i}^\top)^\top$. To estimate $C_{\nu,1,i}$, we use the OLS estimator given by $\hat{\nu}_{1,i} = (\sum_i \mathbf{1}_i^{\chi} \hat{\beta}_{3,i}^{\top} \hat{\beta}_{3,i})^{-1} \sum_i \mathbf{1}_i^{\chi} \hat{\beta}_{3,i}^{\top} \hat{\beta}_{1,i}$. We estimates the weights with $\hat{w}_i = \mathbf{1}_i^{\chi} \left(\text{diag}\left[\hat{v}_i\right]\right)^{-1}$.

To study the asymptotic properties of the estimator $\hat{\nu}$, we consider the following assumption on the size of the cross-section n.

ASSUMPTION A.7: *The size of the cross-section is such that $n = \mathcal{O}(T^{\bar{\gamma}})$ for $\bar{\gamma} > 0$.*

Assumption A.7 puts a bound on the growth of the cross-section such that it does not grow faster that some power of the sample size T. In Proposition 1, we provide the consistency result for the estimator $\hat{\nu}$.

PROPOSITION 1: *(Consistency of $\hat{\nu}$)*
Under Assumptions APR.1 to APR.4, SC.1 and SC.2, B.1 of GOS and Assumptions A.1, A.2, A.4 to A.7, and B.1 to B.5, we have that $\|\hat{\nu} - \nu\| = o_p\left(1\right)$, when $n, T \to \infty$.

Assumptions B.1 to B.5 are discussed in Appendix A. This asymptotic property of $\hat{\nu}$ is studied under the double asymptotics $n, T \to \infty$ in GOS. They show consistency of $\hat{\nu}$ under a full representation of β_i, while we assume a sparse representation of β_i. Hence, our result differs in that respect.

Let us now recover the sparse structure of the conditional expectation of the factors under Assumption A.2. For that purpose, we consider the aLASSO estimator of Zou (2006) to select and estimate the matrix F of coefficients. We solve the following minimization problem for all factor $f_{k,t}, k = 1, \ldots, K$, such that the estimator of the k-th row of the matrix F is given by:

$$(\hat{F}_{0,k}, \hat{F}_{1,k}) = \underset{(F_{0,k}, F_{1,k}) \in \mathbb{R}^{\tilde{p}}}{\text{argmin}} \sum_t \left(f_{k,t} - F_{0,k} - F_{1,k} Z_{t-1}\right)^2 + \delta \sum_{j=1}^{p} \hat{w}_j |F_{1,k,j}|, \quad (10)$$

where δ accounts for the overall amount of shrinkage as in (8), and $\hat{w}_j$ are data dependent weights. Typically, the weights are defined as $\hat{w}_j = {}^{1}\!/\!_{|\hat{F}_{1,k,j}^{\text{OLS}}|^{\check{\gamma}}}$ for $\check{\gamma} > 0$, where $\hat{F}_{1,k,j}^{\text{OLS}}$ are the OLS estimates of $F_{1,k,j}$, the true values in the vector parameter $F_{1,k}$. The estimate $\hat{F}$ stacks row-wise the elements of $(\hat{F}_{0,k}, \hat{F}_{1,k})$ obtained from (10). Under Assumption A.3, no amount of shrinkage is applied to F_0 in F, to always keep the time-invariant contribution in the model. We get the final estimates of the sparse matrix Λ from the relationship $\text{vec}[\hat{\Lambda}^{\top}] = \hat{\nu} + \text{vec}[\hat{F}^{\top}]$, which yields $\hat{\lambda}_t = \hat{\Lambda} Z_{t-1}$. To derive the asymptotic consistency of $\hat{\Lambda}$, we rely on Proposition 1 for the estimator $\hat{\nu}$. Let us consider the following assumption:

ASSUMPTION A.8: *We have $\text{plim}_{T \to \infty} {}^{1}\!/\!_{T} \sum_{t=1}^{T} \tilde{Z}_{t-1} \tilde{Z}_{t-1}^{\top} = \mathbb{E}[\tilde{Z}_{t-1} \tilde{Z}_{t-1}^{\top}]$, where $\mathbb{E}[\tilde{Z}_{t-1} \tilde{Z}_{t-1}^{\top}]$ is a positive definite matrix.*

Assumptions A.8 is a standard regularity assumption on the design matrix for linear regression model, in order to obtain a unique solution for $(F_{0,k}, F_{1,k})$. Under the above Assumption A.8, and Proposition 1, the following proposition gives the consistency result for the estimator $\hat{\Lambda}$.

PROPOSITION 2: *(Consistency of $\hat{\Lambda}$)*

Under Assumptions APR.1 to APR.4, SC.1 and SC.2, B.1 of GOS, Assumptions A.1, A.2, A.4 to A.8 and B.1 to B.6, we have that $\|\hat{\Lambda} - \Lambda\| = o_p(1)$ *, when* $n, T \to \infty$.

Proof of Proposition 2 is direct since from the definition of $\hat{\Lambda}$, $\| \operatorname{vec}[\hat{\Lambda}^\top - \Lambda^\top]\| \leq \|\hat{\nu} - \nu\| + \| \operatorname{vec}[\hat{F}^\top - F^\top]\|$. From Proposition 1, we know that $\|\hat{\nu} - \nu\| = o_p(1)$. Moreover, the aLASSO estimator in (10) is a special case of the estimator in (8), where each group is a singleton. Hence, considering Assumptions A.8 and B.6 which are the counterpart of Assumptions A.4 and A.5 respectively, we get that the result of Lemma 1 applies to (10). Hence, $\| \operatorname{vec}[\hat{F}^\top - F^\top]\| = o_p(1)$. Therefore, we get consistency of $\hat{\lambda}_t$, $\sup_t \|\hat{\lambda}_t - \lambda_t\| = o_p(1)$, under Assumptions A.8 and B.6.

4 Simulation study

In this section, we study how the selection and estimation procedures of Section 3 perform in finite samples. This first simulation study aims at investigating the prediction and selection performance of the aOGL method and at comparing it with the aLASSO method in a very sparse environment (Assumptions A.1 and A.2). To that purpose, we simulate 500 replicates from the DGP in (4) for a (randomly drawn) single asset i with sample size $T_i = 500$. We split that full sample in a training subsample and a testing subsample of 450 and 50 observations. The testing set is used for out-of-sample prediction performance assessment, where we compare the realized excess returns $R_{i,t}$ with their predictions $\hat{R}_{i,t} = \hat{b}_{i,t}^\top \hat{\lambda}_t$ under the model estimated on the training set. Errors in (4) are i.i.d. such that $\varepsilon_{i,t} \sim \mathcal{N}(0, \sigma^2)$, where $\sigma = 0.09$. We match the model specification described in our empirical study (Section 5.1) for the common instruments $Z_{t-1} \in \mathbb{R}^6$ and stock-specific instruments $Z_{i,t-1} \in \mathbb{R}^{13}$. For the factors, we use the Fama-French five-factor model (Fama and French, 2015) described in the next section, namely we condition w.r.t. the values f_t observed in our empirical study for the five factors. We also condition w.r.t. the observed Z_{t-1} and $Z_{i,t}$ for asset i of our empirical study. We only draw the error terms as in a parametric bootstrap.

In accordance with sparsity in Assumptions A.1 and A.2 and non-sparse time-invariant contribution in Assumption A.3, we set the matrices A_i, B_i, and C_i according to their values for asset i in the empirical study, with one non-zero element in B_i and two non-zero element for C_i. We keep the vector A_i full. We set the corresponding $a_{i,t}$ in order to avoid *ex-ante* arbitrage. Since we take very sparse matrices B_i and C_i, we can view the simulation study as conservative for selection performance assessment (type of worst-case scenario). It is in line with the estimation outcome for some stocks in our empirical application. The resulting β_i has 28 non-zero coefficients (including the 6 coefficients induced by the non-sparse time-invariant contribution) over a total of 219 coefficients. The matrices F and Λ are simply set to zero since they do not concern the aOGL estimator.

The selection and prediction performance is measured through the average Root Mean Squared Prediction Error (Av(RMSPE$_R$)), the average Root Mean Squared Error for parameter β_i (Av(RMSE$_\beta$)), the proportion of times the model introduces arbitrage (Arb. (%)), the average number of selected true non-zero coefficients (True+), and average number of regressors in the selected model (NbReg). Table 2 summarizes the results. The aOGL method makes a better job at predicting out-of-sample with a reduction of 1.7% w.r.t. the aLASSO method. The improvement in the average of RMSE for β_i is 109%. The standard errors are also much lower (reduction of 9.1% and 91.0% for the Av(RMSPE$_R$) and Av(RMSE$_\beta$)). Contrary to the aLASSO method, for which 98.2% of estimated models exhibit arbitrage, the aOGL method selects only models without introducing *ex-ante* arbitrage by construction. Since we face less than 100% for the aLASSO method, it sometimes shrinks adequately to zero the coefficients that should be. The large percentage of models with *ex-ante* arbitrage delivered by the aLASSO method is explained by the difficulty of learning the no-arbitrage restrictions from the finite sample information when we do not provide the grouping structure. The aOGL method is able to recover in average the 11 true non-zero coefficients (11.26) while the aLASSO method struggles (7.37). The aOGL method is also more parsimonious than the aLASSO method in terms of selected regressors (average of 14.75 versus 16.05).

Method	Av(RMSPE$_R$)	Av(RMSE$_\beta$)	Arb. (%)	True+	NbReg
aOGL	$9.60 \cdot 10^{-2}$	$1.48 \cdot 10^{-3}$	0.0	11.26	14.75
	$(4.38 \cdot 10^{-4})$	$(9.33 \cdot 10^{-6})$	(-)	(0.20)	(0.31)
aLASSO	$9.76 \cdot 10^{-2}$	$3.10 \cdot 10^{-3}$	98.2	7.37	16.05
	$(4.82 \cdot 10^{-4})$	$(1.04 \cdot 10^{-4})$	(1.12)	(0.10)	(0.61)

Table 2: Performance of estimation and model selection criteria. The methods include the aOGL and aLASSO. We simulate 500 samples under the true sparse DGP. We report the average Root Mean Squared Prediction Error (Av(RMSPE$_R$)), the average Root Mean Squared Error for parameter β_i (Av(RMSE$_\beta$)), the proportion of times the model does not introduce arbitrage (Arb. (%)), the average number of selected true non-zero coefficients (True+), and the average number of regressors in the selected model (NbReg), with their respective standard errors in parenthesis.

Our second simulation set-up focuses on the out-of-sample prediction performance of the aOGL method in a setting close to our empirical study of Section 5.3. We use a training sample to estimate the model and a testing sample to gauge its out-of-sample prediction performance on an equally-weighted portfolio. We consider the same model specification in terms of f_t, Z_{t-1} and $Z_{i,t-1}$ as in the first study and implement the following procedure. We sample randomly a subset of $n = 500$ assets from Section 5 (training sample), while keeping the same proportion of time-invariant models as in Table 4. From each asset i in this subset, we simulate T_i observations from $R_{i,t} = a_{i,t} + b_{i,t}^\top f_t + \varepsilon_{i,t}$ with the coefficients $a_{i,t}$ and $b_{i,t}$ chosen as their aOGL corresponding values for stock i. The 500×1 error vector ε_t at date t is Gaussian with mean zero and block-diagonal correlation matrix with 10 blocks of equal size 50, where, within each block matrix, the correlation between $\varepsilon_{k,t}$ and $\varepsilon_{l,t}$ is set to corr$(\varepsilon_{k,t}, \varepsilon_{l,t}) = 0.25^{|k-l|}$, $k, l = 1, ..., 50, l \neq k$. The variance of each error $\varepsilon_{i,t}$ is set equal to 0.05. From those 500 simulated paths, we implement the aOGL estimation procedure of Section 3, and compare it with the same procedure, but using the aLASSO estimator instead of the aOGL estimator to select the covariates in (5). To evaluate the out-of-sample prediction performance, we simulate one new cross-sectional sample (testing sample) from the

Methods	Av(RMSPE)	Av(MAPE)
aOGL	$6.39 \cdot 10^{-2}$	$4.19 \cdot 10^{-2}$
	$(2.18 \cdot 10^{-3})$	$(7.01 \cdot 10^{-4})$
aLASSO	$9.89 \cdot 10^{-2}$	$4.87 \cdot 10^{-2}$
	$(6.90 \cdot 10^{-3})$	$(4.30 \cdot 10^{-3})$

Table 3: Out-of-sample prediction performance of an equally-weighted portfolio. We compare the aOGL and aLASSO methods. We simulate excess return paths for 500 assets under sparse DGPs. We report the average of Root Mean Squared Prediction Error (Av(RMSPE)), and the average of Mean Absolute Prediction Error (Av(MAPE)) of an equally-weighted portfolio with their respective standard errors in parenthesis.

time-varying factor model for the 500 assets and each date t and compute the prediction $\hat{R}_{i,t} = \hat{b}_{i,t}^{\top}\hat{\lambda}_t$ for the 500 stocks and each date t based on the estimator computed before through the aOGL and aLASSO methods. We finally compute the out-of-sample Prediction Error (PE) for an equally-weighted portfolio through the difference between the new simulated $\frac{1}{500}\sum_i R_{i,t}$ and its predicted value $\frac{1}{500}\sum_i \hat{R}_{i,t}$. We compute the Root Mean Squared Prediction Error (RMSPE), and the Mean Absolute Prediction Error (MAPE) over the vector gathering the PE at each out-of-sample date. We repeat this procedure 100 times to get an average and to compute a standard error. They are reported in Table 3. We can see that the aOGL method is much better at out-of-sample predicting excess returns of an equally-weighted portfolio both in terms of average of MAPE (reduction by 14%) but also in terms of variability as measured by the standard errors (reduction by 84%). The empirical distribution of the prediction errors is given in Figure 1. We can see that the aOGL method is centered closer to zero and with a lower dispersion when compared to the aLASSO method. Those second simulation results again point in favor of our advocated estimation method.

5 Empirical results

This section investigates the predictive capacity of the aOGL estimator and compares it with the aLASSO estimator. We also consider a pure time-invariant model, and a (hybrid) model with constant ν and time-varying risk premia. We use the aLASSO estimator to gauge the added value of incorporating the no-arbitrage restrictions in the penalisation approach and the time-invariant models to gauge the added value of allowing for full time-variation. The latter comparison checks that, when it comes to return prediction, the complicated model does not necessarily outperform because of potential overfitting.

5.1 Data description

We extract the stock returns from the CRSP database for US common stocks listed on the NYSE, AMEX, and NASDAQ, and remove stocks with prices below 5 USD. We exclude financial firms (Standard Industrial Classification Codes between 6000 and

6999). The firm characteristics come from COMPUSTAT. The sample begins in July 1963 and ends in December 2019. It gives us $T = 678$ monthly observations. We proxy the risk-free rate with the 1-month T-bill rate.

From Freyberger et al. (2020), we consider the following $q = 13$ firm level characteristics $Z_{i,t-1}$: change in share outstanding (Δ shrout), log change in the split adjusted shares outstanding (Δ so), growth rate in total assets (Inv), size (LME), last month volume over shares outstanding (lturnover), adjusted profit margin (PM), momentum and intermediate momentum ($r_{12,2}$ and $r_{12,7}$), short-term reversal ($r_{2,1}$), closeness to 52-week high (Rel_to_high), the ratio of market value of equity plus long-term debt minus total assets to Cash and Short-Term Investments (ROC), standard unexplained volume (SUV), and total volume (Tot_vol). We refer to Freyberger et al. (2020) for a detailed description of those characteristics. We only retain stocks for which all 13 characteristics are non-missing. It produces a sample of $n = 6874$. For each $Z_{i,t-1}$, we follow Freyberger et al. (2020) and compute the cross-sectional rank at each time $t-1$ for all observations (see also Chaieb et al., 2021). For the common instruments Z_{t-1}, we consider the $p = 6$ following variables: dividend yield (dp), net equity expansion (ntis), inflation (infl), stock variance (svar), default spread (def_spread), and the term-spread (term_spread). For each Z_{t-1}, we center and standardize all observations.

We consider the two following sets of factors f_t. The first set is the four-factor model of Carhart (1997), such that $f_t = (f_{m,t}, f_{hml,t}, f_{smb,t}, f_{mom,t})^\top$, where $f_{m,t}$ is the month t market excess return over the risk free rate, $f_{hml,t}$, $f_{smb,t}$, $f_{mom,t}$ are respectively the month t returns on zero investment factor-mimicking portfolio for size, book-to-market, and momentum. Our second set of factors considers the profitability factor $f_{rmw,t}$ and the investment factor $f_{cma,t}$ as in the five-factor model of Fama and French (2015), such that $f_t = (f_{m,t}, f_{hml,t}, f_{smb,t}, f_{rmw,t}, f_{cma,t})^\top$. Our choice for a parsimonious specification in the factor space is justified by our goal of studying the selection of common and idiosyncratic instruments Z_{t-1} and $Z_{i,t-1}$ that have impacts on the dynamics of the $a_{i,t}$, $b_{i,t}$, and λ_t. Gagliardini et al. (2019) and Gagliardini et al. (2020) also report evidence that those factors with time-varying loadings are rich enough to achieve a weak cross-sectional dependence in the error terms, namely there are no remaining omitted factors in the error terms.

5.2 In-sample prediction performance and selection results

In this section, we investigate the selection results from the first-pass penalized regression. We compare the fit of the penalized two-pass procedure with aOGL described in Section 3 to the aLASSO estimator, where we select the $x_{i,t}$ and estimate their coefficients in the first-pass regression with the aLASSO estimator of Zou (2006) and fit the WLS estimator for the ν described in Section 2. We compute the estimator $\hat{F}$ as in (10). The horse race starts from the same set of initial data described in the previous section, and the comparison is thus made on the same initial full information. From the characteristics and common instruments outlined in Section 5.1, under the Carhart four-factor model, we have $d = 5$ for the time-invariant model and $d = 199$ for the time-varying model. Regarding the five-factor model of Fama and French (2015), we have $d = 6$ and $d = 219$ for the unconditional and conditional specifications. The number of possible models under the aLASSO method is 2^{194} (2^{213}) with $K = 4$

($K = 5$), while the number of possible models under the aOGL method is 2^{97} (2^{116}), which gives the ratio 2^{-97}, a much lower value than the upper bound $1/8$ in (9).

We choose the regularisation parameter in a data dependent way for each stock to minimize the Akaike Information Criterion (AIC) for both aOGL and aLASSO estimator. As advocated in Greene (2008), we use $\chi_{1,T} = 15$, and require at least 5 years of data such that $\chi_{2,T} = 678/60$. Because of the trimming, we do not keep the same set of stocks for each method and each model. Indeed, due to the different models induced by the first pass for each stock i, the trimming device $\mathbf{1}\{CN(\hat{Q}_{\ddot{x},i}) \leq \chi_{1,T}\}$, yields a different set of stocks for each method. Since we do not wish to introduce multicolinearity in the second-pass regression, we choose to stick with different sets for each method. For the aOGL estimator, the aLASSO estimator, and the time-invariant estimator, we end up with 4412, 2225, 4879 for the four-factor model, and 4441, 2097, 4879 for the five-factor model. We can observe that the trimming device for the aLASSO method is more binding. As seen in the simulation results in Section 4 and in Table 4, the aLASSO method tends to include more variables, and, as a consequence, increase its associated condition number. Table 4 reports the percentage (TI (%)) of estimated models shrunk towards the time-invariant models. For those estimates, we only select the single group corresponding to Restriction R.1 related to Assumption A.3. Around two thirds of the stocks require dynamics in their factor loadings. This new empirical result based on a penalization approach illustrates the relevance of allowing for potential time-variation in modelling excess returns of individual stocks with factor models. Table 4 also reports the percentage (Arb. (%)) of estimated models with time-varying loadings and presenting arbitrage, namely selecting covariates violating the no-arbitrage restrictions. For that computation, both the time-invariant estimates and aOGL estimates avoid *ex-ante* arbitrage by construction. In line with our Monte Carlo results, the aLASSO procedure ends up with all the time-varying models estimated with arbitrage for both factor specifications. We conclude that the aOGL estimation achieves parsimony while avoiding arbitrage in time-varying factor models.

	Carhart four-factor			Fama-French five-factor		
Methods	TI (%)	Arb. (%)	Av NbReg	TI (%)	Arb. (%)	Av NbReg
aOGL	38	0	13.24	35	0	14.15
aLASSO	36	100	33.45	31	100	37.20
time-invariant	100	0	5	100	0	6

Table 4: Percentage (TI (%)) of estimated models shrunk towards the time-invariant specification, percentage (Arb. (%)) of estimated time-varying models presenting arbitrage and average number of regressors selected (Av NbReg) with the Carhart four-factor and Fama-French five-factor models for the aOGL, aLASSO, and time-invariant methods. The sample of US equity excess returns begins in July 1963 and ends in December 2019.

In the three first lines of Tables 5 and 6, we investigate the type of stock excess returns that exhibit time-variation in the their factor loadings. For both factor specifications, the longer the sample size, the more "action" is needed for the dynamics of the

Carhart four-factor						
T_i	$\leq$ 6y	6y - 10y	20y - 30y	30y - 40y	40y - 50y	$\geq$ 50y
Nber of stocks	480	1535	1542	921	393	406
Av. # of sel. var.	10.89	10.80	11.45	13.21	14.32	28.81
TI (%)	42.08	44.36	38.59	32.25	27.23	23.65
dp (%)	13.12	12.83	15.89	25.52	41.22	59.36
ntis (%)	16.04	17.79	27.89	40.61	46.06	40.89
infl (%)	27.71	26.58	25.88	32.25	39.44	40.39
svar (%)	21.04	18.76	18.94	15.20	6.87	30.79
def_spread (%)	22.08	23.71	31.97	34.96	42.24	49.51
term_spread (%)	34.58	31.73	29.83	35.94	31.04	27.59
Δ shrout (%)	0.21	0.26	0.39	0.54	0.25	10.34
Δ so (%)	0.00	0.13	0.32	0.98	0.51	10.34
Inv (%)	0.00	0.39	0.32	0.54	0.51	7.64
LME (%)	0.63	0.33	0.26	0.87	0.25	6.65
lturnover (%)	0.83	0.52	0.39	0.98	0.25	7.64
PM (%)	0.21	0.20	0.26	0.54	0.51	6.65
$r_{12,2}$ (%)	0.63	0.59	0.26	0.87	0.51	10.59
$r_{12,7}$ (%)	0.21	0.20	0.26	0.11	0.00	10.10
$r_{2,1}$ (%)	0.00	0.13	0.19	0.76	0.51	7.39
Rel_to_high (%)	0.00	0.39	0.32	0.22	0.25	7.39
ROC (%)	0.63	0.33	0.19	0.76	0.25	6.16
SUV (%)	0.63	0.26	0.13	0.65	0.00	7.14
Tot_vol (%)	0.00	0.13	0.26	0.54	0.51	6.40

Table 5: Selection results sorted by sample size (T_i) for the Carhart four-factor specification. We first report the number of stocks (Nber of stocks), the average number of selected variables (Av. # of sel. var.) and the percentage of estimated models shrunk towards the time-invariant specification (TI (%)) per sample size range. Then we give the percentage w.r.t. the total number of stocks of each of the 6 variables in Z_{t-1} and the 13 variables in $Z_{i,t-1}$ per stock excess return sample size. The sample of US equity excess returns begins in July 1963 and ends in December 2019.

	Fama-French five-factor					
T_i	$\leq$ 6y	6y - 10y	20y - 30y	30y - 40y	40y - 50y	$\geq$ 50y
Nber of stocks	480	1535	1542	921	393	406
mean # of sel. var.	12.71	12.46	12.88	14.14	15.09	24.76
TI (%)	39.79	40.59	36.12	32.46	30.79	19.70
dp (%)	15.00	13.29	15.30	22.80	32.32	49.26
ntis (%)	18.12	20.91	29.57	39.31	45.04	51.97
infl (%)	31.46	30.81	26.07	33.33	36.64	48.28
svar (%)	24.79	21.50	18.74	15.20	6.87	24.63
def_spread (%)	25.62	25.86	32.49	36.26	40.20	46.80
term_spread (%)	35.83	34.40	32.04	36.48	34.61	40.64
Δ shrout (%)	0.42	0.33	0.65	0.98	0.51	4.93
Δ so (%)	0.42	0.20	0.13	0.76	0.51	5.67
Inv (%)	0.42	0.33	0.52	0.43	0.25	4.68
LME (%)	0.42	0.39	0.32	0.98	0.51	4.19
lturnover (%)	0.63	0.39	0.58	1.09	0.51	0.23
PM (%)	0.00	0.00	0.65	0.65	0.51	3.69
$r_{12,2}$ (%)	1.04	0.46	0.84	0.76	0.25	5.17
$r_{12,7}$ (%)	0.21	0.20	0.52	0.65	0.25	4.43
$r_{2,1}$ (%)	0.42	0.20	0.06	0.43	0.25	4.19
Rel_to_high (%)	0.42	0.33	0.52	0.33	0.25	4.68
ROC (%)	0.42	0.33	0.32	0.98	0.51	3.94
SUV (%)	0.42	0.26	0.39	0.98	0.25	5.17
Tot_vol (%)	0.00	0.00	0.58	0.54	0.51	3.69

Table 6: Selection results sorted by sample size (T_i) for the Fama-French five-factor specification. We first report the number of stocks (Nber of stocks), the average number of selected variables (Av. # of sel. var.) and the percentage of estimated models shrunk towards the time-invariant specification (TI (%)) per sample size range. Then, we give the percentage w.r.t. the total number of stocks of each of the 6 variables in Z_{t-1} and the 13 variables in $Z_{i,t-1}$ per stock excess return sample size. The sample of US equity excess returns begins in July 1963 and ends in December 2019.

	Carhart four-factor			
	f_m	f_{hml}	f_{smb}	f_{mom}
dp (%)	18.44	14.66	15.41	14.59
ntis (%)	23.19	19.28	17.59	22.12
infl (%)	20.68	20.59	17.27	24.82
svar (%)	12.67	10.23	9.22	12.48
def_spread (%)	24.43	20.78	19.93	24.82
term_spread (%)	23.52	20.94	20.52	25.70
Δ shrout (%)	0.88	0.75	0.68	0.78
Δ so (%)	1.07	1.14	0.94	1.11
Inv (%)	0.75	0.62	0.65	0.65
LME (%)	0.62	0.42	0.46	0.39
lturnover (%)	1.11	0.85	0.88	0.85
PM (%)	0.59	1.10	0.62	0.55
$r_{12,2}$ (%)	1.17	0.85	0.85	0.85
$r_{12,7}$ (%)	1.01	0.88	0.85	2.73
$r_{2,1}$ (%)	0.98	0.62	0.59	0.98
Rel_to_high (%)	0.75	0.85	1.51	1.37
ROC (%)	0.88	0.82	0.75	0.91
SUV (%)	0.94	0.78	0.75	0.88
Tot_vol (%)	0.81	0.72	0.65	0.72

Table 7: Selection results for the Carhart four-factor specification. For stocks exhibiting time variation in their factor loadings, we report the percentage of each of the 6 variables in Z_{t-1} and the 13 variables in $Z_{i,t-1}$ selected per factor. The sample of US equity excess returns begins in July 1963 and ends in December 2019.

	Carhart four-factor			
	f_m	f_{hml}	f_{smb}	f_{mom}
dp		✓	✓	✓
ntis	✓	✓	✓	✓
infl	✓	✓		
svar	✓	✓	✓	✓
def_spread	✓	✓	✓	✓
term_spread	✓	✓		✓

Table 8: Selection results for the drivers of $\mathbb{E}[f_t|\mathcal{F}_{t-1}]$ for the Carhart four-factor specification. A check denotes inclusion of a covariate in Z_{t-1}. The sample begins in July 1963 and ends in December 2019.

	Fama-French five-factor				
	f_m	f_{hml}	f_{smb}	f_{rmw}	f_{cma}
dp (%)	16.45	12.49	13.63	8.48	8.95
ntis (%)	23.43	17.87	18.34	13.14	13.09
infl (%)	22.71	18.41	17.84	15.08	14.20
svar (%)	13.73	7.56	9.65	5.21	7.05
def_spread (%)	24.95	19.51	20.34	12.83	15.15
term_spread (%)	26.88	19.51	21.47	14.90	15.18
Δ shrout (%)	0.66	0.51	0.44	0.40	0.32
Δ so (%)	0.73	0.51	0.60	0.40	0.47
Inv (%)	0.32	0.44	0.35	0.23	0.32
LME (%)	0.32	0.16	0.22	0.05	0.19
lturnover (%)	0.44	0.35	0.33	0.25	0.28
PM (%)	0.47	0.37	0.33	0.44	0.32
$r_{12,2}$ (%)	0.89	0.54	0.51	0.48	0.38
$r_{12,7}$ (%)	0.63	0.51	0.54	0.44	0.47
$r_{2,1}$ (%)	0.47	0.35	0.38	0.51	0.19
Rel_to_high (%)	0.60	0.57	0.57	0.66	0.41
ROC (%)	0.73	0.40	0.66	0.37	0.19
SUV (%)	0.85	0.48	0.47	0.55	0.41
Tot_vol (%)	0.51	0.22	0.40	0.33	0.51

Table 9: Selection results for the Fama-French five-factor specification. For stocks exhibiting time variation in their factor loadings, we report the percentage of each of the 6 variables in Z_{t-1} and the 13 variables in $Z_{i,t-1}$ selected per factor. The sample of US equity excess returns begins in July 1963 and ends in December 2019.

	Fama-French five-factor				
	f_m	f_{hml}	f_{smb}	f_{rmw}	f_{cma}
dp		✓	✓	✓	✓
ntis	✓	✓	✓	✓	✓
infl	✓	✓			
svar	✓	✓	✓		
def_spread	✓	✓	✓	✓	✓
term_spread	✓	✓			

Table 10: Selection results for the drivers of $\mathbb{E}[f_t|\mathcal{F}_{t-1}]$ for the Fama-French five-factor specification. A check denotes inclusion of a covariate in Z_{t-1}. The sample begins in July 1963 and ends in December 2019.

factor loadings. Indeed, the aOGL method selects a time-invariant model for more than 50% of the stocks excess returns exhibiting historical data smaller than 10 years, and this for both factor specifications. On the contrary, 80% of the models with the longest sample size ($\geq$ 50 years) need time-variation in their factor loadings.

In the next lines of Tables 5 and 6, we further look at the selected variables among the 6 in Z_{t-1} and the 13 in $Z_{i,t-1}$. Across all sample sizes T_i, the percentages of selected variables in Z_{t-1} are much higher than the percentages of selected variables in $Z_{i,t-1}$. For smaller time spans, some characteristics are never selected. Therefore, it seems that the common instruments Z_{t-1} are key drivers of the time variation of the factor loadings. It is particularly true for the range $\geq$ 50y. It shows the need of including common instruments that pick up the influence of the business cycles on the factor loading dynamics in larger time spans. For those sample sizes, the higher selection rate of Z_{t-1} and $Z_{i,t-1}$ is not automatically due to the choice of tuning parameters, but to the higher complexity of the dynamics of factor loading. In long time spans, we need to capture the heterogeneity of the different states of the economy and of the firms through time by using additional Z_{t-1} and $Z_{i,t-1}$. Further investigation shows that stock-specific instruments $Z_{i,t-1}$ are more often selected for small-cap stocks, while common instruments Z_{t-1} are more often selected for large-cap stocks. Large-cap stocks are more homogeneous and so firm characteristics matter less than common instruments. While there are no common instruments that are never selected, and this across all time spans, the proportions far below 100% demonstrate the need to select instruments in a data-driven way.

In Tables 7 and 9, we report the percentage that the 6 variables in Z_{t-1} (scaled factors) and the 13 variables in $Z_{i,t-1}$ are selected through the aOGL method for both factor specifications. The percentages of selected common instruments are similar for the factors f_m, f_{hml}, and f_{smb} shared between the two models. With the Carhart four-factor specification, we need more variables in $Z_{i,t-1}$ to describe the dynamics of the factor loadings in comparison with the Fama-French five-factor specification. In line with Chaieb et al. (2021), the characteristics are not necessarily paired more often with their corresponding factors. The size characteristic LME is more often paired with the market factor f_m than with the size factor f_{smb}. On the contrary, the momentum characteristics $r_{12,7}$ and $r_{2,1}$ are often associated with its corresponding factor f_{mom}. Finally, Tables 8 and 10 show that the conditional expectations of the factors f_{rmw} and f_{cma} in the Fama-French five-factor specification need less covariates than for the other factors. The variables ntis and def_spread are selected for all factors.

Let us now investigate in-sample predictability performance. As, in Chaieb et al. (2021), we decompose the conditional expected return of asset i for month t for both time-varying factor specifications, as:

$$\mathbb{E}\left[R_{i,t}|\mathcal{F}_{t-1}\right] = a_{i,t} - b_{i,t}^{\top}\nu_t + b_{i,t}^{\top}\lambda_t = a_{i,t} + b_{i,t}^{\top}\mathbb{E}[f_t|\mathcal{F}_{t-1}]. \tag{11}$$

For such time-varying specifications, the contribution of the pricing errors $a_{i,t} - b_{i,t}^{\top}\nu_t$ is often small, revealing that the no-arbitrage restrictions are met for a vast majority of dates. When they are not, Chaieb et al. (2021) show that incorporating pricing errors, instead of only relying on $b_{i,t}^{\top}\lambda_t$ in (11), helps to predict future equity excess returns. Similarly, for the time-invariant models, we decompose the unconditional expected return as:

$$\mathbb{E}\left[R_{i,t}\right] = a_i - b_i^\top \nu + b_i^\top \lambda = a_i + b_i^\top \mathbb{E}[f_t]. \tag{12}$$

For such time-invariant specifications, the contribution of the pricing errors $a_i - b_i^\top \nu$ is often large. We also consider the case of constant ν and time-varying risk premia λ_t ($\lambda_t \& \nu$), for which we decompose the conditional expected return as

$$\mathbb{E}\left[R_{i,t}|\mathcal{F}_{t-1}\right] = a_i - b_i^\top \nu + b_i^\top \lambda_t = a_i + b_i^\top \mathbb{E}[f_t|\mathcal{F}_{t-1}]. \tag{13}$$

In such a hybrid model (Avramov, 2004), the time-variation in $\mathbb{E}[R_{i,t}|\mathcal{F}_{t-1}]$ only comes from the time-variation in $\mathbb{E}[f_t|\mathcal{F}_{t-1}]$ since ν is constant because of the no-arbitrage restrictions with constant b_i and a_i.

To compare the prediction performance of the four estimation approaches, we compute the RMSPE of an equally-weighted portfolio for the Carhart four-factor model and Fama-French five-factor model. Equal weighting corresponds to cross-sectional averaging. Chaieb et al. (2021) also uses this weighting scheme. For that portfolio, we compute the PE by comparing the prediction made at time t by each model ((11), (12) and 13) to the forward 12-months realized excess returns, namely the average of the realized excess returns over the next 12 months. Table 11 reports the RMSPE, as well as the Av(|PE|) and Std(|PE|) for the Carhart four-factor model and Fama-French five-factor model specifications. The aOGL method performs better than its natural competitor, the aLASSO, even for that very diversified stable portfolio, where we expect differences in prediction performance to be attenuated. It is comparable in terms of the RMSPE to the $\lambda_t \& \nu$ method, with a lower Std(|PE|). Figure 2 displays the corresponding boxplots of the PE computed at each month for each method. The boxplots for the aOGL method in Figure 2 are narrower than for the aLASSO method, and comparable for the two other methods. Those predictability improvements against the aLASSO approach provide further evidence in support for the aOGL approach advocated for the first-pass regression, so that we can incorporate model parameter restrictions to get models compatible *ex-ante* with the no-arbitrage restrictions. To further investigate time-varying predictability, Figures 4 to 7 show the forward 12-months realized excess returns for the equally-weighted portfolio and compare them with the predicted excess returns computed from (11) and (12) for the two methods with penalisation, respectively for the Carhart four-factor and Fama-French five-factor specifications. In both Figures 4 and 6, the aOGL predicted excess return paths (red plain line) overall reconcile well with the realized excess returns (black dashed line). On the contrary, the aLASSO method in Figures 5 and 7 does not reconcile well the predicted excess returns with the realized excess returns and sometimes predicts large negative excess returns, which is at odd with a positive reward expected from taking risks. The observed differences in the decomposition between estimates of $a_{i,t}$ (orange shaded area) and of $b_{i,t}^\top \mathbb{E}[f_t|\mathcal{F}_{t-1}]$ (blue shaded area) come from the selected regressors in the first pass. Since the aLASSO penalization ends up with time-varying models presenting arbitrage, we observe larger values for estimated $\hat{a}_{i,t}$, especially during the recession periods (gray areas) determined by the National Bureau of Economic Research (NBER). The aOGL method avoids putting covariates in estimated $\hat{a}_{i,t}$ that should not be there because of the no-arbitrage restrictions. Besides, the estimated path for $a_{i,t}$ is close to zero with the aOGL method as it should be if we believe that the factors are most of the time fully tradable.

	Carhart four-factor			Fama-French five-factor		
Methods	RMSPE	Av(|PE|)	Std(|PE|)	RMSPE	Av(|PE|)	Std(|PE|)
aOGL	$1.46 \cdot 10^{-2}$	$1.12 \cdot 10^{-2}$	$0.93 \cdot 10^{-2}$	$1.49 \cdot 10^{-2}$	$1.16 \cdot 10^{-2}$	$0.93 \cdot 10^{-2}$
aLASSO	$1.62 \cdot 10^{-2}$	$1.27 \cdot 10^{-2}$	$1.01 \cdot 10^{-2}$	$2.14 \cdot 10^{-2}$	$1.67 \cdot 10^{-2}$	$1.35 \cdot 10^{-2}$
TI	$1.79 \cdot 10^{-2}$	$1.36 \cdot 10^{-2}$	$1.18 \cdot 10^{-2}$	$1.37 \cdot 10^{-2}$	$1.02 \cdot 10^{-2}$	$0.91 \cdot 10^{-2}$
$\lambda_t \& \nu$	$1.45 \cdot 10^{-2}$	$1.08 \cdot 10^{-2}$	$0.97 \cdot 10^{-2}$	$1.46 \cdot 10^{-2}$	$1.08 \cdot 10^{-2}$	$0.98 \cdot 10^{-2}$

Table 11: Root Mean Squared Prediction Error (RMSPE), Mean Absolute Prediction Error (Av(|PE|)) and Standard Deviation of the Absolute Prediction Error (Std(|PE|)) of an equally-weighted portfolio with the Carhart four-factor and Fama-French five-factor models for the aOGL, aLASSO, time-invariant (TI) and $\lambda_t \& \nu$ methods. The sample of US equity excess returns begins in July 1963 and ends in December 2019.

5.3 Out-of-sample prediction performance

In this section, we compare the out-of-sample prediction performance for the same methods used in the previous section. Here, we compute PE but for data that never enter into model estimation. We follow a similar approach to Gu et al. (2020). We split the sample into two subsamples, one for training and one for testing. We estimate the models from July 1963 to December 2009 and compute PE from January 2010 to December 2019 (recent period). We repeat the same analysis for a training period from July 1963 to December 1999 and a testing period from January 2000 to December 2009 (older period). We closely follow the same setting as in the previous section, the only difference being that we separate the subsample used for estimation from the one used for prediction performance assessment.

	Carhart four-factor					
	Jan. 2000 to Dec. 2009			Jan. 2010 to Dec. 2019		
Methods	RMSPE	Av(|PE|)	Std(|PE|)	RMSPE	Av(|PE|)	Std(|PE|)
aOGL	$1.58 \cdot 10^{-2}$	$1.23 \cdot 10^{-2}$	$1.00 \cdot 10^{-2}$	$1.34 \cdot 10^{-2}$	$1.06 \cdot 10^{-2}$	$0.83 \cdot 10^{-2}$
aLASSO	$2.43 \cdot 10^{-2}$	$2.03 \cdot 10^{-2}$	$1.37 \cdot 10^{-2}$	$7.44 \cdot 10^{-2}$	$6.17 \cdot 10^{-2}$	$4.18 \cdot 10^{-2}$
TI	$1.70 \cdot 10^{-2}$	$1.32 \cdot 10^{-2}$	$1.08 \cdot 10^{-2}$	$1.70 \cdot 10^{-2}$	$1.32 \cdot 10^{-2}$	$1.08 \cdot 10^{-2}$
$\lambda_t \& \nu$	$1.57 \cdot 10^{-2}$	$1.24 \cdot 10^{-2}$	$0.96 \cdot 10^{-2}$	$1.79 \cdot 10^{-2}$	$1.31 \cdot 10^{-2}$	$1.22 \cdot 10^{-2}$

Table 12: Out-of-sample Root Mean Squared Prediction Error (RMSPE), Mean Absolute Prediction Error (Av(|PE|)) and Standard Deviation of the Absolute Prediction Error (Std(|PE|)) of an equally-weighted portfolio with the Carhart four-factor model for the aOGL, aLASSO, time-invariant (TI) and $\lambda_t \& \nu$ methods. The testing periods are Jan. 2000 to Dec. 2009 and Jan. 2010 to Dec. 2019. Their associated training periods precede them and start in July 1963.

	Fama-French five-factor					
	Jan. 2000 to Dec. 2009			Jan. 2010 to Dec. 2019		
Methods	RMSPE	Av(|PE|)	Std(|PE|)	RMSPE	Av(|PE|)	Std(|PE|)
aOGL	$1.86 \cdot 10^{-2}$	$1.38 \cdot 10^{-2}$	$1.24 \cdot 10^{-2}$	$1.26 \cdot 10^{-2}$	$0.99 \cdot 10^{-2}$	$0.77 \cdot 10^{-2}$
aLASSO	$9.05 \cdot 10^{-2}$	$5.11 \cdot 10^{-2}$	$7.49 \cdot 10^{-2}$	$6.63 \cdot 10^{-2}$	$5.99 \cdot 10^{-2}$	$2.86 \cdot 10^{-2}$
TI	$1.70 \cdot 10^{-2}$	$1.32 \cdot 10^{-2}$	$1.07 \cdot 10^{-2}$	$1.69 \cdot 10^{-2}$	$1.32 \cdot 10^{-2}$	$1.07 \cdot 10^{-2}$
$\lambda_t\&\nu$	$1.56 \cdot 10^{-2}$	$1.24 \cdot 10^{-2}$	$0.96 \cdot 10^{-2}$	$1.79 \cdot 10^{-2}$	$1.32 \cdot 10^{-2}$	$1.21 \cdot 10^{-2}$

Table 13: Out-of-sample Root Mean Squared Prediction Error (RMSPE), Mean Absolute Prediction Error (Av(|PE|)) and Standard Deviation of the Absolute Prediction Error (Std(|PE|)) of an equally-weighted portfolio with the Fama-French five-factor model for the aOGL, aLASSO, time-invariant (TI) and $\lambda_t\&\nu$ methods. The testing periods are Jan. 2000 to Dec. 2009 and Jan. 2010 to Dec. 2019. Their associated training periods precede them and start in July 1963.

	Carhart four-factor							
	Jan. 2000 to Dec. 2009				Jan. 2010 to Dec. 2019			
Year	aOGL	aLASSO	TI	$\lambda_t\&\nu$	aOGL	aLASSO	TI	$\lambda_t\&\nu$
1	0.72	0.46	0.87	0.89	0.72	0.35	0.87	0.89
2	0.50	0.41	0.84	0.83	0.71	0.28	0.85	0.92
3	0.20	0.32	0.66	0.70	0.74	0.40	0.67	0.53
4	0.37	0.41	0.60	0.65	0.63	0.37	0.61	0.66
5	0.41	0.42	0.61	0.64	0.63	0.32	0.62	0.68
6	0.40	0.40	0.23	0.40	0.57	0.27	0.23	0.12
7	0.38	0.40	0.18	0.35	0.60	0.08	0.18	0.05
8	0.23	0.19	0.24	0.38	0.62	-0.81	0.24	0.13
9	-0.19	-1.85	0.18	0.35	0.61	-0.86	0.18	0.07

Table 14: Out-of-sample R^2 of an equally-weighted portfolio with the Carhart four-factor model for the aOGL, aLASSO, time-invariant (TI) and $\lambda_t\&\nu$ methods. The out-of-sample R^2 are computed for each year of the testing periods from Jan. 2000 to Dec. 2009 and from Jan. 2010 to Dec. 2019. Their associated training periods precede them and start in July 1963.

We see that the aOGL method performs better than the aLASSO method in all cases as shown in Tables 12 to 13. Furthermore, the aOGL method often performs better than a time-invariant method as exhibited by the RMSPE and the lower Std(|PE|). Such an advantage over time-invariant alternatives is less clear for the out-of-sample R^2 computed each year on the whole testing periods in Tables 14 and 15. We follow Gu et al. (2020) (see also Gu et al. (2021)), and compute the out-of-sample R^2 as: $R^2 = 1 - \frac{\sum_{i,t}(R_{i,t}-\hat{R}_{i,t})^2}{\sum_{i,t} R_{i,t}^2}$, where the $R_{i,t}$ and $\hat{R}_{i,t}$ are the out-of-sample observed and predicted returns of each stock in the equally-weighted portfolio for the dates in each testing period.

On the contrary, the aOGL method keeps a strong advantage over the aLASSO

	Fama-French five-factor							
	Jan. 2000 to Dec. 2009				Jan. 2010 to Dec. 2019			
Year	aOGL	aLASSO	TI	$\lambda_t\&\nu$	aOGL	aLASSO	TI	$\lambda_t\&\nu$
1	0.69	0.32	0.86	0.89	0.72	0.25	0.87	0.87
2	0.50	0.31	0.84	0.83	0.74	0.20	0.84	0.91
3	0.16	0.17	0.66	0.70	0.79	0.07	0.66	0.50
4	0.35	0.19	0.60	0.65	0.67	-0.06	0.60	0.65
5	0.39	0.22	0.61	0.65	0.65	-0.20	0.61	0.68
6	0.38	0.23	0.23	0.40	0.60	-0.29	0.23	0.11
7	0.35	0.23	0.18	0.35	0.64	-0.57	0.18	0.05
8	0.30	0.38	0.24	0.38	0.60	-0.19	0.24	0.13
9	0.09	0.30	0.18	0.34	0.58	-1.17	0.18	0.07

Table 15: Out-of-sample R^2 of an equally-weighted portfolio with the Fama-French five-factor model for the aOGL, aLASSO, time-invariant (TI) and $\lambda_t\&\nu$ methods. The out-of-sample R^2 arew computed for each year of the testing periods from Jan. 2000 to Dec. 2009 and from Jan. 2010 to Dec. 2019. Their associated training periods precede them and start in July 1963.

method, especially for the years closer to the training periods. The deterioration in terms of prediction performance at the end of the first testing period is explained by the aftermath of the 2008 financial crisis. The good prediction performance at the beginning of the second testing period is explained by incorporating the 2008 financial crisis in the estimation sample. The reported out-of-sample R^2 are in the same range as the ones given by Gu et al. (2021) for managed portfolios. For both testing periods, the boxplots in Figure 3 show that out-of-sample PE related to the portfolio excess returns for the aOGL method are located closer to zero, more symmetrically distributed, and narrower. As observed in the in-sample analysis, the aOGL method seems to perform better in terms of out-of-sample predictability as shown by the distributional behavior of the PE. We believe that the good out-of-sample performance for the portfolio comes from the diversification of the prediction errors among the single assets. We observe a similar phenomenon in forecast combinations (Timmermann, 2006).

We provide a decomposition of the importance of each variable for the out-of-sample prediction performance in Figure 8. We delete one Z_{t-1} at a time in the two testing periods for both factor specifications. A negative difference shows a deterioration of predictability measured by the out-of-sample R^2 on the whole testing period when a particular Z_{t-1} is taken out when forming out-of-sample predictions. The 95% confidence intervals built by a percentile block bootstrap approach give the information on whether the difference is significantly different from zero at a 5% significance level. We see that the confidence intervals are wide in the first testing period. The most important Z_{t-1} in terms of out-of-sample predictability is term_spread, and then we have infl and ntis for the Carhart four-factor model. The most important Z_{t-1} in terms of out-of-sample predictability is ntis, and then we have def_spread and infl for the Fama-French five-factor model. In the second testing period, we observe only a slight deterioration, the largest one being for term_spread, albeit not significant for all variables with narrow confidence intervals.

6 Conclusions

Our empirical results show that taking explicitly into account the no-arbitrage restrictions coming from the Arbitrage Pricing Theory do help in predictive modeling of large cross-sectional equity data sets with penalisation methods. We view this approach as an example of a structural approach to big data where incorporating finance theory improves on the prediction performance of the estimated quantities. It resonates with structural approaches in panel econometrics guided by economic theory (Bonhomme and Shaikh, 2017). In asset management and risk management, a better predictive performance of excess returns should help to better gauge time-variation in the risk-reward trade-off. In asset selection, it should help to improve performance of time-varying portfolio allocation when we use predicted excess returns as inputs. From our simulation and empirical results, we expect our procedure to perform well in out-of-sample prediction for portfolio building.

7 Acknowledgements

We would like to thank the Co-Editor T. Andersen, Associate Editor, and two referees for constructive criticism and numerous suggestions which have led to substantial improvements over the previous version. We are grateful to I. Chaieb, T. Chordia, A.-P. Fortin, P. Gagliardini, R. Garcia, M. Karemera, H. Langlois, M. Medeiros, A. Patton, S. Pruitt, J. Thimme, S. van den Hoff, D. Xiu, P. Zaffaroni for their helpful comments, as well as seminar participants at Queen Mary, University of Nottingham, Louvain Finance Seminar, IAE Lille, Université d'Orléans, LUISS, Auburn University Statistics and Data Science Seminar, CREST-ENSAE Financial Econometrics Seminar, and participants at the 8^{th} days of Econometrics for Finance, 2021 SoFiE Machine Learning conference, Vienna Workshop "Econometrics of Option Markets", 2021 SFI Research days, 2021 North America Summer Meeting of the Econometric Society, 13^{th} Annual SoFiE Conference, 2021 IAAE Conference, 2021 EcoSta, 7^{th} IYFS Conference, 2021 China International Conference in Finance (CICF), 2021 EEA-ESEM, 20th Workshop in Econometrics for Finance, 2021 CFE-CM Statistics, 19th Paris December Finance Meeting, ML approaches Finance and Management workshop, 38^{th} AFFI conference, QFFE conference 2022, Frontiers of Factor Investing conference 2022, and Webseminar AI & Big Data in Finance Research Forum. G. Bakalli and O. Scaillet were supported by the SNSF Grant $\#100018-182582$. S. Guerrier was supported by the SNSF Professorships Grant #176843 and by the Innosuisse-Boomerang Grant #37308.1 IP-ENG.

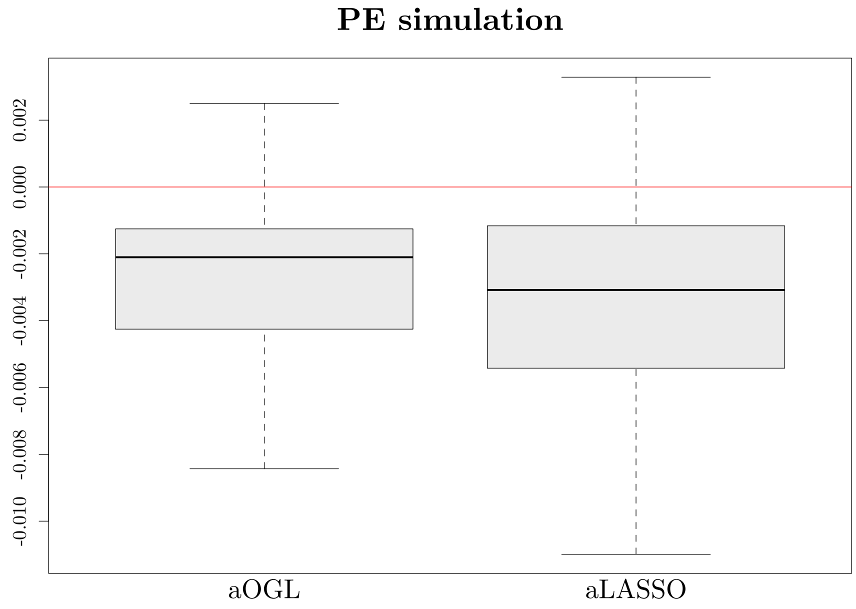

Figure 1: Empirical distribution of out-of-sample Prediction Error (PE) of an equally-weighted portfolio. We compare the aOGL and aLASSO methods. We simulate excess return paths for 500 assets under sparse DGPs.

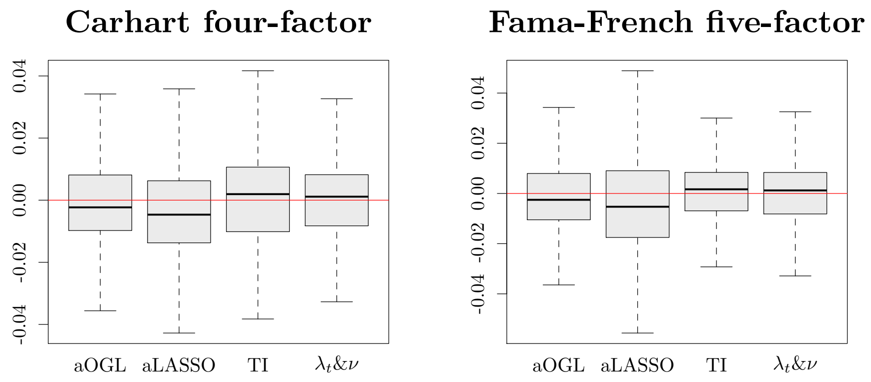

Figure 2: Empirical distribution of in-sample Prediction Error (PE) of an equally-weighted portfolio. We compare the aOGL, aLASSO, time-invariant (TI) and $\lambda_t\&\nu$ methods. The left panel corresponds to the Carhart four-factor model. The right panel corresponds to the Fama-French five-factor model. The sample of US equity excess returns begins in July 1963 and ends in December 2019.

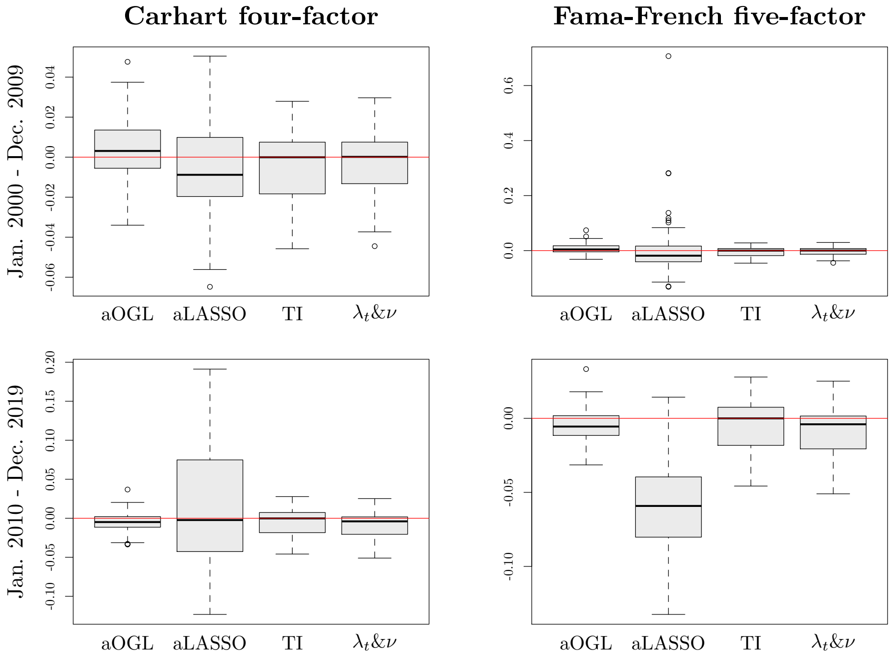

Figure 3: Empirical distribution of out-of-sample Prediction Error (PE) of an equally-weighted portfolio. We compare the aOGL, aLASSO, time-invariant (TI) and $\lambda_t\&\nu$ methods for the Carhart four-factor and Fama-French five-factor models. The upper panels are for the testing period 2000-2009. The lower panels are for the testing period 2010-2019. Their associated training periods precede them and start in July 1963.

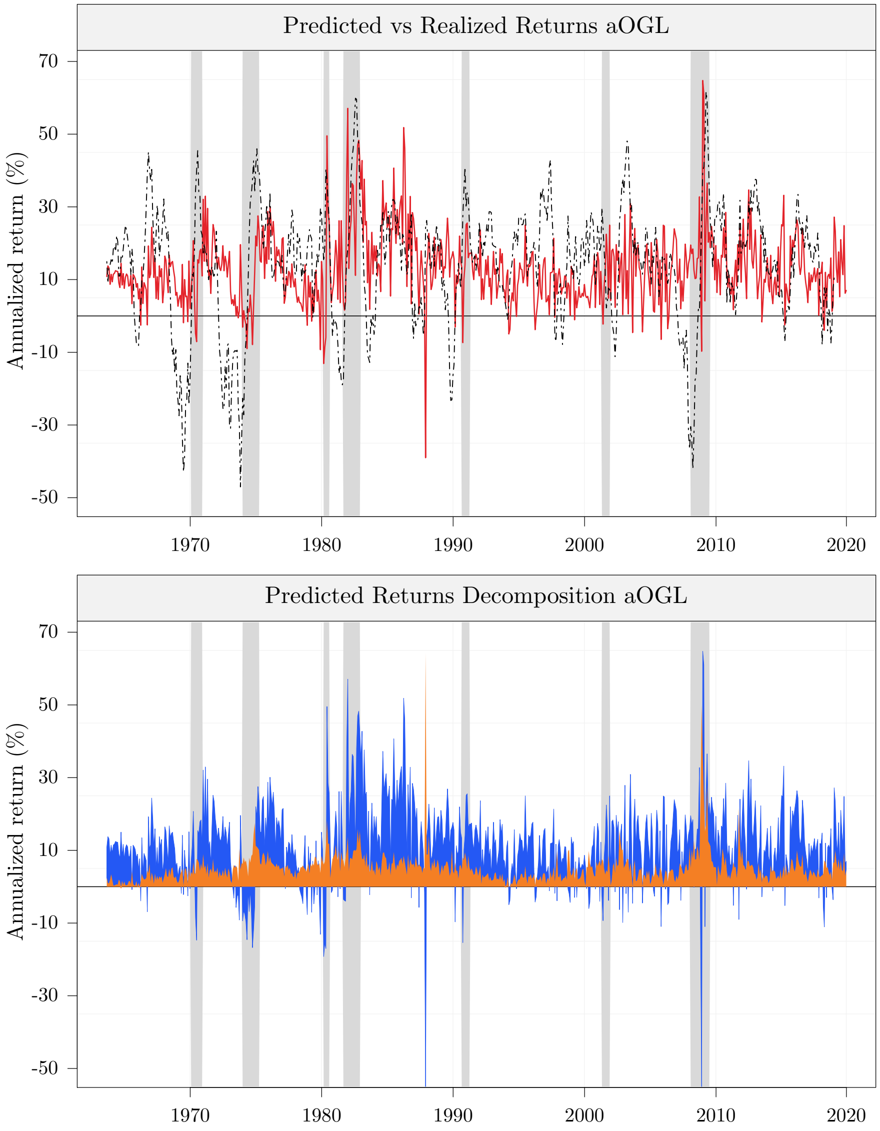

Figure 4: Predicted excess returns, realized excess returns, and prediction decomposition for the Carhart four-factor model and an equally-weighted portfolio with the aOGL method. In the upper panel, the predicted excess return path corresponds to the red plain line. The realized excess returns correspond to the black dashed line. In the lower panel, the orange shaded area corresponds to estimates of $a_{i,t}$. The blue shaded area corresponds to estimates of $b_{i,t}^{\top}\mathbb{E}[f_t|\mathcal{F}_{t-1}]$. The gray shaded areas correspond to the recession periods determined by the National Bureau of Economic Research (NBER). The sample of US equity excess returns begins in July 1963 and ends in December 2019.

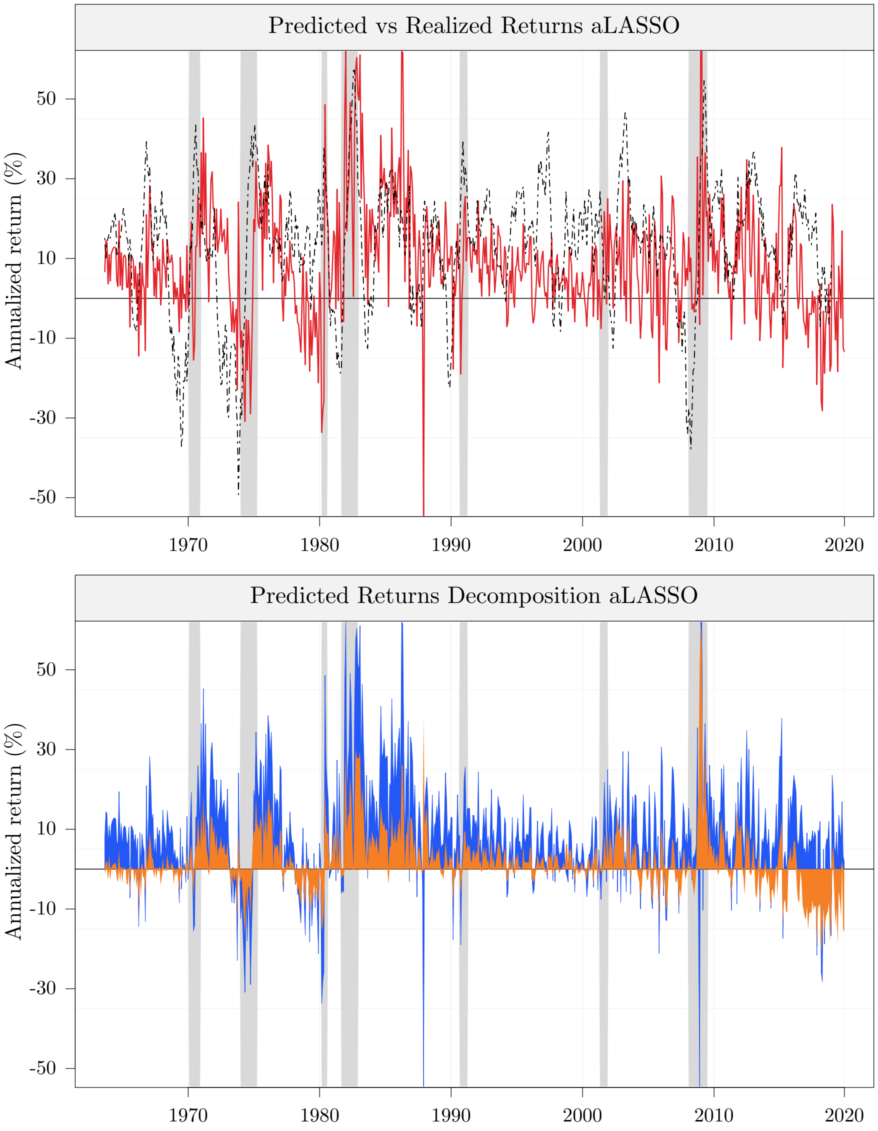

Figure 5: Predicted excess returns, realized excess returns, and prediction decomposition for the Carhart four-factor model and an equally-weighted portfolio with the aLASSO method. In the upper panel, the predicted excess return path corresponds to the red plain line. The realized excess returns correspond to the black dashed line. In the lower panel, the orange shaded area corresponds to estimates of $a_{i,t}$. The blue shaded area corresponds to estimates of $b_{i,t}^{\top}\mathbb{E}[f_t|\mathcal{F}_{t-1}]$. The gray shaded areas correspond to the recession periods determined by the National Bureau of Economic Research (NBER). The sample of US equity excess returns begins in July 1963 and ends in December 2019.

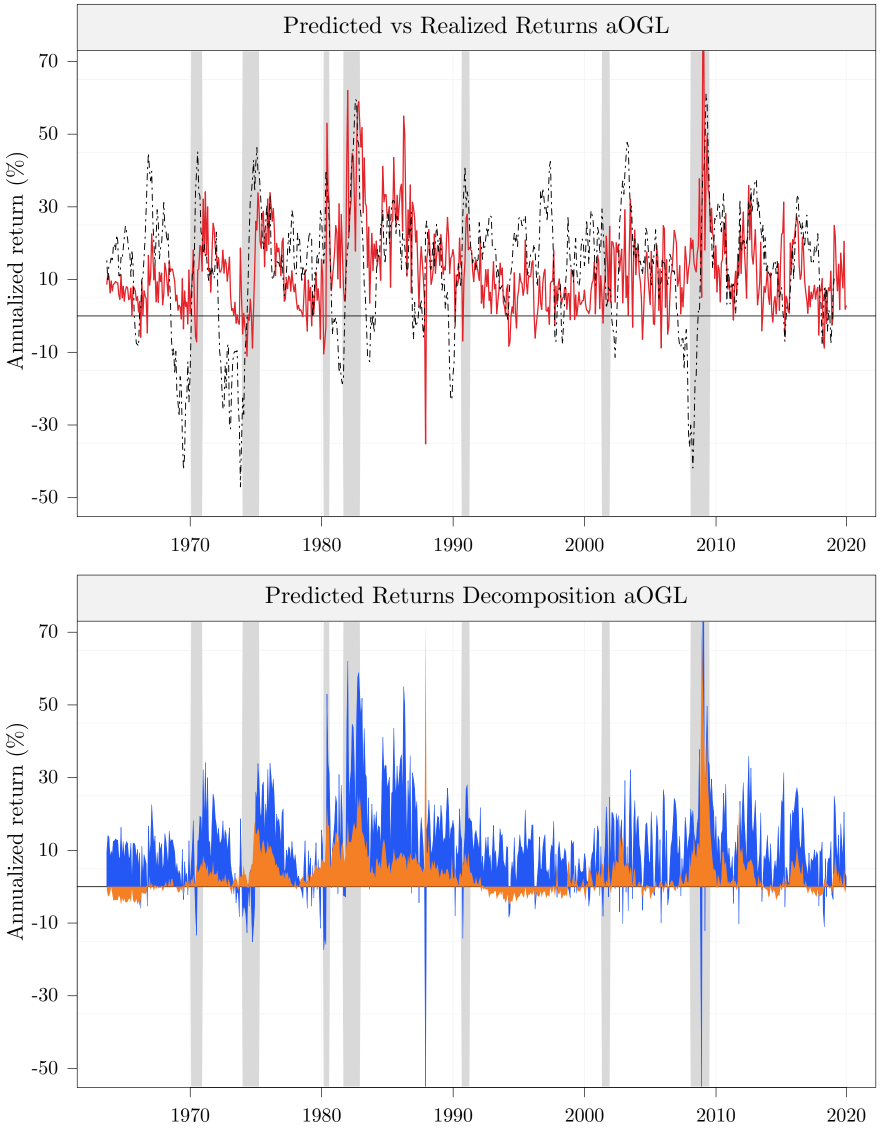

Figure 6: Predicted excess returns, realized excess returns, and prediction decomposition for the Fama-French five-factor model and an equally-weighted portfolio with the aOGL method. In the upper panel, the predicted excess return path corresponds to the red plain line. The realized excess returns correspond to the black dashed line. In the lower panel, the orange shaded area corresponds to estimates of $a_{i,t}$. The blue shaded area corresponds to estimates of $b_{i,t}^{\top}\mathbb{E}[f_t|\mathcal{F}_{t-1}]$. The gray shaded areas correspond to the recession periods determined by the National Bureau of Economic Research (NBER). The sample of US equity excess returns begins in July 1963 and ends in December 2019.

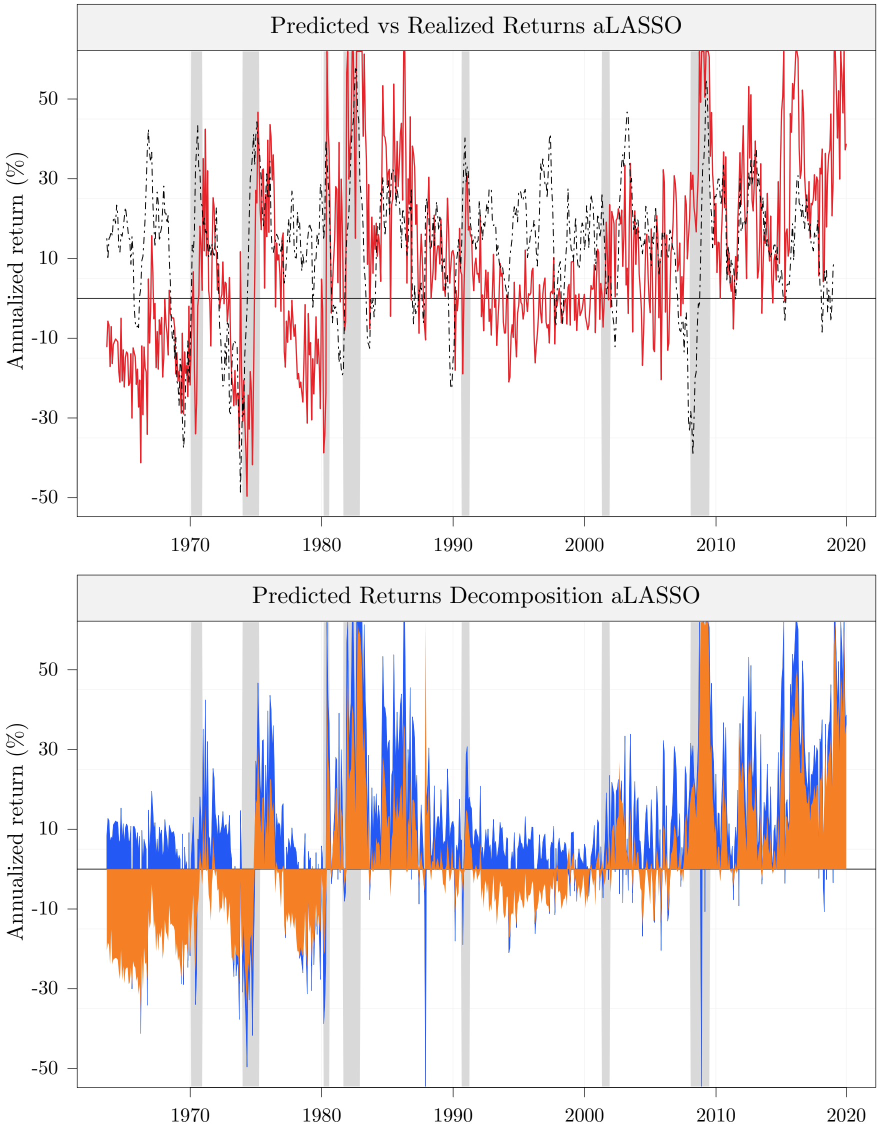

Figure 7: Predicted excess returns, realized excess returns, and prediction decomposition for the Fama-French five-factor model and an equally-weighted portfolio with the aLASSO method. In the upper panel, the predicted excess return path corresponds to the red plain line. The realized excess returns correspond to the black dashed line. In the lower panel, the orange shaded area corresponds to estimates of $a_{i,t}$. The blue shaded area corresponds to estimates of $b_{i,t}^{\top}\mathbb{E}[f_t|\mathcal{F}_{t-1}]$. The gray shaded areas correspond to the recession periods determined by the National Bureau of Economic Research (NBER). The sample of US equity excess returns begins in July 1963 and ends in December 2019.

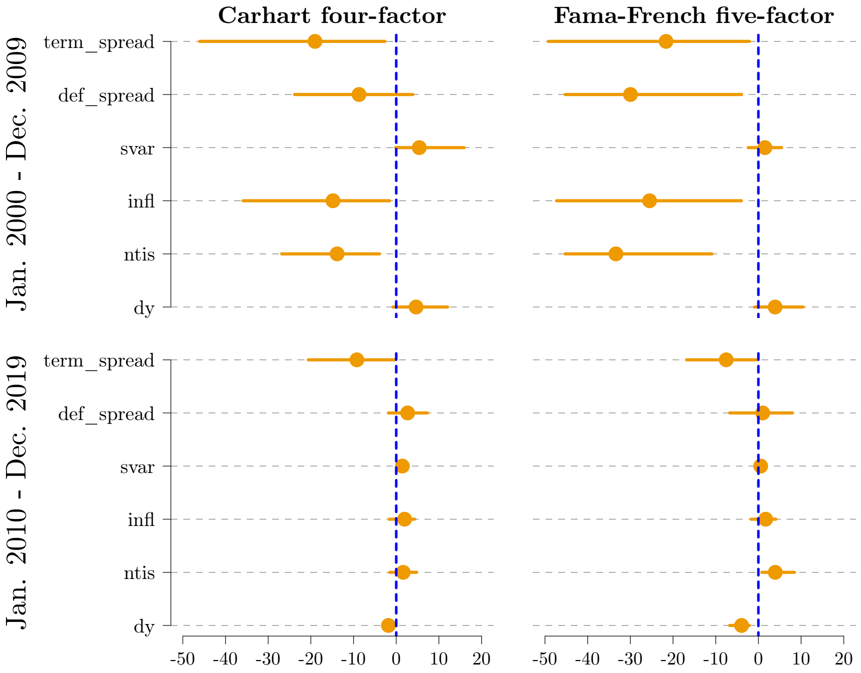

Figure 8: Out-of-sample variable importance for the Cahart four-factor model and Fama-French five-factor model. We compute the difference in out-of-sample R^2 between the full model and a model when one Z_{t-1} at a time is removed from the full model. The orange dots show the point estimates while the orange lines show the 95% confidence intervals computed by a percentile block bootstrap approach. The upper panel presents the results for the Jan. 2000 - Dec. 2009 testing period, while the lower panel shows the results for the Jan. 2010 - Dec. 2019 testing period.

References

Aït-Sahalia, Y., Jacod, J., and Xiu, D. (2020). Inference on risk premia in continuous-time asset pricing models. Technical report, National Bureau of Economic Research.

Aït-Sahalia, Y. and Xiu, D. (2017). Using principal component analysis to estimate a high dimensional factor model with high-frequency data. Journal of Econometrics, 201(2):384–399.

Al-Najjar, N. (1995). Decomposition and characterization of risk with a continuum of random variables. Econometrica, 63(5):1195–1224.

Avramov, D. (2004). Stock return predictability and asset pricing models. The Review of Financial Studies, 17(3):699–738.

Avramov, D., Cheng, S., Metzker, L., and Voigt, S. (2022). Integrating factor models. Journal of Finance, forthcoming.

Avramov, D. and Chordia, T. (2006). Asset pricing models and financial market anomalies. Review of Financial Studies, 19(3):1000–1040.

Bach, F. R. (2008). Consistency of the group lasso and multiple kernel learning. Journal of Machine Learning Research, 9(6).

Black, F., Jensen, M., and Scholes, M. (1972). The Capital Asset Pricing Model: Some empirical findings. In Jensen, M., editor, Studies in the Theory of Capital Markets. Praeger Publishers Inc.

Bonhomme, S. and Shaikh, A. M. (2017). Keeping the ECON in Econometrics: (micro-) econometrics in the Journal of Political Economy. Journal of Political Economy, 125(6):1846–1853.

Bryzgalova, S. (2015). Spurious factors in linear asset pricing models. Technical report, London School of Economics.

Bryzgalova, S., Huang, J., and Julliard, C. (2019). Bayesian solutions for the factor zoo: We just ran two quadrillion models. Journal of Finance, forthcoming.

Carhart, M. M. (1997). On persistence in mutual fund performance. Journal of Finance, 52(1):57–82.

Chaieb, I., Langlois, H., and Scaillet, O. (2021). Factors and risk premia in individual international stock returns. Journal of Financial Economics, 141(2):669–692.

Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. Econometrica, 51(5):1281–1304.

Chen, L., Pelger, M., and Zhu, J. (2022). Deep learning in asset pricing. Management Science, forthcoming.

Chernozhukov, V., Chetverikov, D., Kato, K., and Koike, Y. (2022). Improved Central Limit Theorem and bootstrap approximations in high dimensions. The Annals of Statistics, forthcoming.

Chinco, A., Clark-Joseph, A. D., and Ye, M. (2019). Sparse signals in the cross-section of returns. The Journal of Finance, 74(1):449–492.

Cochrane, J. H. (1996). A cross-sectional test of an investment-based asset pricing model. Journal of Political Economy, 104(3):572–621.

Cochrane, J. H. (2011). Presidential address: Discount rates. The Journal of Finance, 66(4):1047–1108.

Cong, W., Feng, G., He, J., and He, X. (2022a). Asset pricing with panel tree under global split criteria. Technical report, Cornell University.

Cong, W., Tang, K., Wang, J., and Zhang, Y. (2022b). Alphaportfolio: Direct construction through deep reinforcement learning and interpretable AI. Technical report, Cornell University.

Das, D. and Lahiri, S. (2021). Central Limit Theorem in high dimensions: The optimal bound on dimension growth rate. Transactions of the American Mathematical Society, 374(10):6991–7009.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. Journal of Financial Economics, 116(1):1–22.

Fama, E. F. and MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. Journal of Political Economy, 81(3):607–636.

Fan, J., Furger, A., and Xiu, D. (2016). Incorporating global industrial classification standard into portfolio allocation: A simple factor-based large covariance matrix estimator with high-frequency data. Journal of Business & Economic Statistics, 34(4):489–503.

Fan, J., Ke, T., Liao, Y., and Neuhierl, A. (2022). Structural deep learning in conditional asset pricing. Technical report, Rutgers University.

Fan, J., Masini, R., and Medeiros, M. (2021). Bridging factor and sparse models. Technical report, Princeton University.

Feng, G., Giglio, S., and Xiu, D. (2020). Taming the factor zoo: A test of new factors. Journal of Finance, 75(3):1327–1370.

Ferson, W. E. and Harvey, C. R. (1991). The variation of economic risk premiums. Journal of Political Economy, 99(2):385–415.

Freyberger, J., Neuhierl, A., and Weber, M. (2020). Dissecting characteristics nonparametrically. Review of Financial Studies, 33(5):2326–2377.

Gagliardini, P., Ossola, E., and Scaillet, O. (2016). Time-varying risk premium in large cross-sectional equity data sets. Econometrica, 84(3):985–1046.

Gagliardini, P., Ossola, E., and Scaillet, O. (2019). A diagnostic criterion for approximate factor structure. Journal of Econometrics, 212(2):503–521.

Gagliardini, P., Ossola, E., and Scaillet, O. (2020). Estimation of large dimensional conditional factor models in finance. In Durlauf, S., Hansen, L. P., Heckman, J. J., and Matzkin, R. L., editors, Handbook of Econometrics, Volume 7A, chapter 3, pages 219–282. North Holland.

Giannone, D., Lenza, M., and Primiceri, G. E. (2021). Economic predictions with big data: The illusion of sparsity. Econometrica, 89(5):2409–2437.

Greene, W. H. (2008). Econometrics Analysis. Upper Saddle River: Prentice Hall.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical asset pricing via machine learning. Review of Financial Studies, 33(5):2223–2273.

Gu, S., Kelly, B., and Xiu, D. (2021). Autoencoder asset pricing models. Journal of Econometrics, 222(1):429–450.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). Statistical learning with sparsity. Monographs on Statistics and Applied Probability, 143:143.

Jacob, L., Obozinski, G., and Vert, J.-P. (2009). Group lasso with overlap and graph lasso. In Proceedings of the 26th Annual International Conference on Machine Learning, pages 433–440.

Jagannathan, R., Skoulakis, G., and Wang, Z. (2010). The analysis of the cross-section of security returns. In Aït-Sahalia, Y. and Hansen, L. P., editors, Handbook of Financial Econometrics: Applications, chapter 13, pages 73–134. North Holland.

Jagannathan, R. and Wang, Z. (1996). The conditional CAPM and the cross-section of expected returns. Journal of Finance, 51(1):3–53.

Jagannathan, R. and Wang, Z. (1998). An asymptotic theory for estimating beta-pricing models using cross-sectional regression. Journal of Finance, 53(4):1285–1309.

Kan, R., Robotti, C., and Shanken, J. (2013). Pricing model performance and the two-pass cross-sectional regression methodology. Journal of Finance, 68(6):2617–2649.

Lopes, M. (2022). Central Limit Theorem and bootstrap approximations in high dimensions: Near $1/\sqrt{n}$ rates via implicit smoothing. The Annals of Statistics, forthcoming.

Lounici, K., Pontil, M., Van De Geer, S., and Tsybakov, A. B. (2011). Oracle inequalities and optimal inference under group sparsity. The Annals of Statistics, 39(4):2164–2204.

Nardi, Y. and Rinaldo, A. (2008). On the asymptotic properties of the group lasso estimator for linear models. Electronic Journal of Statistics, 2:605–633.

Pelger, M. and Xiong, R. (2019). State-varying factor models of large dimensions. Journal of Business & Economic Statistics, forthcoming.

Percival, D. (2012). Theoretical properties of the overlapping groups lasso. Electronic Journal of Statistics, 6:269–288.

Ross, S. (1976). The arbitrage theory of capital asset pricing. Journal of Economic Theory, 13(3):341–360.

Shanken, J. (1985). Multivariate tests of the zero-beta capm. Journal of Financial Economics, 14(3):327–348.

Shanken, J. (1990). Intertemporal asset pricing: An empirical investigation. Journal of Econometrics, 45(1-2):99–120.

Shanken, J. (1992). On the estimation of beta-pricing models. Review of Financial Studies, 5(1):1–33.

Shanken, J. and Zhou, G. (2007). Estimating and testing beta pricing models: Alternative methods and their performance in simulations. Journal of Financial Economics, 84(1):40–86.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1):267–288.

Timmermann, A. (2006). Forecast combinations. In Elliott, G., Granger, C., and Timmermann, A., editors, Handbook of Economic Forecasting, chapter 4, pages 135–196. North Holland.

van der Vaart, A. W. and Wellner, J. A. (1998). Weak Convergence and Empirical Processes: With Applications to Statistics. Springer.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(1):49–67.

Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American Statistical Association, 101(476):1418–1429.

A Regularity conditions

This Appendix lists and comments the regularity conditions needed to derive the asymptotic properties of the estimation procedure (see also Appendix A in GOS). Beforehand, recall the following vector $x_{i,t} = (\mathrm{vech}[X_t]^\top, \tilde{Z}_{t-1}^\top \otimes Z_{i,t-1}^\top, f_t^\top \otimes \tilde{Z}_{t-1}^\top, f_t^\top \otimes Z_{i,t-1}^\top)^\top$ of dimension d.

ASSUMPTION B.1: *There exists a constant M such that a) $\sup_i \|x_{i,t}\| \leq M$, P-a.s.. Moreover, b) $\sup_i \|A_i\| < \infty$, $\sup_i \|B_i\| < \infty$, $\sup_i \|C_i\| < \infty$.*

ASSUMPTION B.2: $\inf_i \mathbb{E}[I_{i,t}|\gamma_i] > 0$.

ASSUMPTION B.3: $\inf_i \mathrm{eig}_{\mathrm{min}}(\mathbb{E}[x_{i,t}x_{i,t}^\top|\gamma_i]) > 0$*, where* $\mathrm{eig}_{\mathrm{min}}$ *denotes the minimum eigenvalue of* $\mathbb{E}[x_{i,t}x_{i,t}^\top|\gamma_i]$.

ASSUMPTION B.4: *The trimming constants satisfy $\chi_{1,T} = \mathcal{O}\left((\log T)^{\kappa_1}\right)$, $\chi_{2,T} = \mathcal{O}\left((\log T)^{\kappa_2}\right)$, with $\kappa_1, \kappa_2 > 0$.*

ASSUMPTION B.5: *For all $n \in \mathbb{N}\backslash\{0\}$, there exist sub-Gaussian random variables $Y_{i,l} \sim \mathrm{subG}(\sigma^2), \sigma^2 < \infty$ such that $\mathbb{E}[\max_{i,l}|\sqrt{T_i}(\hat{\beta}_{i,l} - \beta_{i,l})|] \leq \mathbb{E}[\max_{i,l}|Y_{i,l}|]$, for $i = 1, \ldots, n$, and $l = 1, \ldots, d$.*

ASSUMPTION B.6: *We have that $\mathbb{E}[u_t|u_{\underline{t-1}}, \mathcal{F}_t] = 0$ and there exists a constant $M > 0$, such that $\|\mathbb{E}[u_t u_t^\top|Z_{t-1}]\| \leq M$, for all t, where $u_t = f_t - \mathbb{E}[f_t|\mathcal{F}_{t-1}]$.*

Assumption B.1 eases the proofs and requires uniform upper bounds on the regressor values, intercept, and model coefficients. Assumption B.2 implies that the fraction of the time period in which an asset return is observed is bounded away from zero asymptotically uniformly across assets, while Assumption B.3 bounds away from zero the minimum eigenvalue of the population squared moment to exclude asymptotic multicolinearity problems uniformly across assets. Assumption B.4 gives an upper bound on the divergence rate of the trimming constants such that logarithmic divergence rate allows to control the aOGL estimation error in the second-pass regression. Assumption B.5 is a technical requirement on $\mathbb{E}[\max_{i,l}|\sqrt{T_i}(\hat{\beta}_{i,l} - \beta_{i,l})|]$. From Lemma 1, we have that $\hat{\beta}_{i,l}$ are asymptotically normally distributed and we might think that Assumption B.5 is directly satisfied due the properties of sub-Gaussian random variables. However, it is not the case due to our double asymptotics with $n, T \to \infty$. To illustrate the necessity of this requirement, we can consider the following example: let $Z_i \sim \mathcal{U}(0,1)$ and $\delta_{i,T} = i\,\mathbf{1}_{i \geq T}$ for $i = 1, \ldots, n$. For all i, we have $X_{i,T} = Z_i + \delta_{i,T} \implies \mathcal{U}(0,1)$ as $T \to \infty$. Suppose that $n > T$, then we have $\lim_{T\to\infty} \mathbb{E}[\max_{i=1,\ldots,n}|Z_i|] \leq 1$ while $\lim_{T\to\infty} \mathbb{E}[\max_{i=1,\ldots,n}|X_{i,T}|]$ diverges. We can replace Assumption B.5 by other requirements, for example, by considering sub-Gaussian error terms in the first-pass regression. Sub-Gaussianity is often used in the literature on inference in high-dimensions; see e.g. Das and Lahiri (2021), Chernozhukov et al. (2022), Lopes (2022). Finally, Assumption B.6 allows for a martingale difference sequence and bounds the conditional variance-covariance matrix for the linear innovation u_t associated with the factor process. This assumption helps to prove consistency of the aLASSO estimator

$\hat{F}_k$ using the same arguments as in Lemma 1.

B Proof of Lemma 1

We follow the proof strategy of Percival (2012) (see Nardi and Rinaldo (2008) for related arguments for the Group-LASSO). Let $\beta_i^\star = \beta_i + \frac{u_i}{\sqrt{T_i}}$ and $\{v_{i,g}^\star\}$ and $\{v_{i,g}\}$ be decomposition of β_i minimizing $\|\beta_i^\star\|_{2,1,\mathcal{G}}$ and $\|\beta_i\|_{2,1,\mathcal{G}}$, respectively. Multiplying (8) by $\frac{T_i}{2}$, we have that

$$\mathcal{Q}^\star(u_i) = \frac{1}{2}\sum_t \left(I_{i,t}R_{i,t} - \left(\beta_i + \frac{u_i}{\sqrt{T_i}}\right)^\top I_{i,t}x_{i,t}\right)^2 + \delta T_i \sum_g \delta_g \left\| v_{i,g} + \frac{1}{\sqrt{T_i}} v_{i,g}^{u_i}\right\|,$$

where $v_{i,g}^{u_i} = \sqrt{T_i}(v_{i,g}^\star - v_{i,g})$ is a decomposition of $u_i = \sqrt{T_i}(\beta_i^\star - \beta_i)$. We define

$$\hat{u}_i = \underset{u_i \in \mathbb{R}^d}{\operatorname{argmin}}\, \mathcal{Q}^\star(u_i),$$

then we have $\hat{\beta}_i = \beta_i + \frac{\hat{u}_i}{\sqrt{T_i}}$. We write $D^\star(u_i) = \mathcal{Q}^\star(u_i) - \mathcal{Q}^\star(0)$ and thus we obtain

$$\begin{aligned} D^\star(u_i) &= \frac{1}{2}u_i^\top \hat{Q}_{x,i}u_i - \frac{1}{\sqrt{T_i}}u_i^\top \sum_t I_{i,t}x_{i,t}\varepsilon_{i,t} \\ &\quad + \sqrt{T_i}\delta \sum_g \delta_g \sqrt{T_i}\left(\left\| v_{i,g} + \frac{1}{\sqrt{T_i}}v_{i,g}^{u_i}\right\| - \|v_{i,g}\|\right) \\ &= \mathcal{I}_1 + \sum_g \mathcal{I}_{2,g}. \end{aligned}$$

From Percival (2012), we know that, for $g \in G_{H_i}, \mathcal{I}_{2,g}$ vanishes to zero since δ_g based on an initial $\sqrt{T_i}$-consistent estimator goes to $\|v_{i,g}\|^{-\tilde{\gamma}}$, from Assumption A.6, the uniqueness of the decomposition of $v_{i,g}$ and $\sqrt{T_i}\delta = o(1)$. Moreover, for $g \in G_{H_i^c}, \mathcal{I}_{2,g}$ diverges and, for $g \in G_{H_{0,i}}, \mathcal{I}_{2,g}$ diverges since $T_i^{\tilde{\gamma}/2}\|v_{i,g}^{\text{init}}\|^{\tilde{\gamma}} = \mathcal{O}_p(1)$ and $T_i^{(1+\tilde{\gamma})/2}\delta$ diverges, where $v_{i,g}^{\text{init}}$ is the initial data dependent estimator of the latent decomposition of β_i. Moreover, under Assumption A.4 and A.5, using the CLT for martingale difference sequences and Slutsky's theorem, we have that

$$\mathcal{I}_1 \Longrightarrow \frac{1}{2}u_i^\top Q_{x,i}u_i - u_i^\top W_i,$$

where $W_i \sim \mathcal{N}(0, \sigma_i^2 Q_{x,i})$. It follows that

$$D^\star(u_i) \Longrightarrow D(u_i),$$

with

$$D(u_i) = \begin{cases} \frac{1}{2}u_i^\top Q_{x,i}u_i - u_i^\top W_i, & \text{if } v_{i,g}^{u_i} \neq 0, \text{ for } g \in G_{H_i}, \\ \infty, & \text{else.} \end{cases}$$

Minimizing $D(u_i)$ and using the argmax theorem from van der Vaart and Wellner (1998) conclude the proof as in Percival (2012).

□

C Proof of Proposition 1

From Lemma 1, we have the convergence in distribution to a Gaussian random variable for all $i = 1, \ldots, n$:

$$\sqrt{T_i}\left(\hat{\beta}_i - \beta_i\right) \Longrightarrow V_i.$$

Next, we consider the expectation of $\sup_i \mathbf{1}_i^{\chi} \|\hat{\beta}_i - \beta_i\|$:

$$\begin{aligned}
\mathbb{E}\left[\sup_{1\le i\le n} \mathbf{1}_i^{\chi}\|\hat{\beta}_i - \beta_i\|\right] &= \mathbb{E}\left[\max_{1\le i\le n} \mathbf{1}_i^{\chi}\sqrt{\sum_{l=1}^{d}\left(\hat{\beta}_{i,l} - \beta_{i,l}\right)^2}\right] \\
&\le \sqrt{d}\,\mathbb{E}\left[\max_{\substack{1\le i\le n\\1\le l\le d}} \mathbf{1}_i^{\chi}\left|\hat{\beta}_{i,l} - \beta_{i,l}\right|\right] \\
&= \frac{\sqrt{d}}{\sqrt{T}}\mathbb{E}\left[\max_{\substack{1\le i\le n\\1\le l\le d}} \mathbf{1}_i^{\chi}\sqrt{\frac{T}{T_i}}\sqrt{T_i}\left|\hat{\beta}_{i,l} - \beta_{i,l}\right|\right] \\
&\le \frac{\sqrt{d\,\chi_{2,T}}}{\sqrt{T}}\mathbb{E}\left[\max_{\substack{1\le i\le n\\1\le l\le d}} \sqrt{T_i}\left|\hat{\beta}_{i,l} - \beta_{i,l}\right|\right].
\end{aligned}$$

From Assumption B.5, we have that

$$\begin{aligned}
\frac{\sqrt{d\,\chi_{2,T}}}{\sqrt{T}}\mathbb{E}\left[\max_{\substack{1\le i\le n\\1\le l\le d}} \sqrt{T_i}\left|\hat{\beta}_{i,l} - \beta_{i,l}\right|\right] &\le \frac{\sqrt{d\,\chi_{2,T}}}{\sqrt{T}}\left(\mathbb{E}\left[\max_{\substack{1\le i\le n\\1\le l\le d}} |Y_{i,l}|\right]\right) \\
&\le \frac{\sqrt{2d\,\chi_{2,T}\,\sigma^2\log(2nd)}}{\sqrt{T}}.
\end{aligned}$$

Thus, by Assumption A.7 and B.4, there exist a positive constant C such that

$$\mathbb{E}\left[\sup_{1\le i\le n} \mathbf{1}_i^{\chi}\|\hat{\beta}_i - \beta_i\|\right] \le C\sqrt{\frac{\log(T)^{1+\kappa_2}}{T}},$$

and we have

$$\lim_{n\to\infty}\sqrt{\frac{T}{\log(T)^{1+\kappa_2}}}\mathbb{E}\left[\sup_{1\le i\le n} \mathbf{1}_i^{\chi}\|\hat{\beta}_i - \beta_i\|\right] \le C.$$

We have by Markov's inequality that, for any $\epsilon > 0$,

$$\begin{aligned}
&\Pr\left(\sqrt{\frac{T}{\log(T)^{1+\kappa_2}}}\sup_{1\le i\le n} \mathbf{1}_i^{\chi}\|\hat{\beta}_i - \beta_i\| \ge \epsilon\right) \\
&\quad\le \sqrt{\frac{T}{\epsilon^2\log(T)^{1+\kappa_2}}}\,\mathbb{E}\left[\sup_{1\le i\le n} \mathbf{1}_i^{\chi}\|\hat{\beta}_i - \beta_i\|\right] \le \frac{C}{\epsilon}.
\end{aligned}$$

Thus, we have

$$\sup_{1\leq i\leq n} \mathbf{1}_i^{\chi} \|\hat{\beta}_i - \beta_i\| = \mathcal{O}_p\left(\sqrt{\frac{\log(T)^{1+\kappa_2}}{T}}\right),$$

implying that

$$\sup_{1\leq i\leq n} \mathbf{1}_i^{\chi} \|\hat{\beta}_i - \beta_i\| = o_p(1).$$

The consistency of $\hat{\nu}$ then follows from the following results ii) $\sup_i \|w_i\| = O(1)$, iii) ${}^1\!/_n \sum_i \|\hat{w}_i - w_i\| = o_p(1)$ and iv) $\hat{Q}_{\beta_3} - Q_{\beta_3} = o_p(1)$ of Lemma 3 of GOS under Assumptions A.4 and B.1 to B.4. As in the proof of Proposition 3 of GOS, they ensure $\|\hat{\nu} - \nu\| = o_p\left(1\right)$, which concludes the proof.

□

D No-arbitrage *ex-ante* grouping structure

This section is dedicated to describe how to construct the grouping structure needed to create the vector $\tilde{x}_{i,t}$ of duplicated regressors from the original $x_{i,t}$. We use the duplicated regressors to implement the numerical optimisation of the aOGL method. From the set of Restrictions R.1 to R.4, it appears that, for any element in $x_{1,i,t}$ related to a specific element of $\tilde{Z}_{t-1,l}$ and $Z_{i,t-1,m}$, there exist multiple corresponding regressors in $x_{2,i,t}$ related to the same instrument l and characteristic m. To implement a shrinkage estimator satisfying Restrictions R.1 to R.4, we define the following sets of indices. The first group related to Restriction R.1 always includes all covariates corresponding to the time-invariant contribution. Hence, we define $\tilde{x}^{(1)}_{i,t} = (x_{i,t,j})_{j \in \iota_{g_1}} \in \mathbb{R}^{n_1}$, where $n_1 = K + 1$, and ι_{g_1} is a set of indices such that,

$$\iota_{g_1} = \{1, d_1 + 1, \ldots, d_1 + k\tilde{p} + 1, \ldots, d_1 + (K-1)\tilde{p} + 1\} \in \mathbb{N}_+^{K+1},$$

for $k = 1, \ldots, K-1$ and with $\mathbb{N}_+ = \mathbb{N} \setminus \{0\}$. The next set of groups are related to Restriction R.2, and we define $\tilde{x}^{(2)}_{i,t} = (x_{i,t,j})_{j \in \iota_{g_2}} \in \mathbb{R}^{n_2}$, where $n_2 = \tilde{p}(\tilde{p}-1)/2$, and the set ι_{g_2} corresponds to the indices related to the non-diagonal elements of $\mathrm{vech}(X_t)$ in $x_{i,t}$. To characterize it, let us first define the set of indices related to the diagonal elements in $\mathrm{vech}(X_t)$ (i.e., the squared elements $Z^2_{t-1,l}$) and the index set related to all elements in $\mathrm{vech}\,(X_t)$ as follows

$$\mathcal{D} = \left\{ x \in \mathbb{N}_+ \mid x = 1 + (k-1)(\tilde{p}+1) - \frac{(k-1)k}{2}, k \in \{1, ..., \tilde{p}\} \right\},$$
$$\mathcal{A} = \left\{ x \in \mathbb{N}_+ | x \leq \frac{(\tilde{p}+1)\tilde{p}}{2} \right\},$$

such that the indices in $\mathcal{A} \backslash \mathcal{D}$ generate the set of indices:

$$\iota_{g_2} = \{\iota_{g_{2,1}}, \ldots, \iota_{g_{2,n_2}}\} \in \mathbb{N}_+^{n_2}.$$

Let us describe the group structure needed within a regular Group-LASSO by replicating our covariates to solve the original aOGL problem and ensuring that Restrictions R.3 and R.4 are met. First, the scalar u_l, for $l = 1, \ldots, p$, denotes the l-th element of the set $\mathcal{D} \setminus \{1\}$, i.e., the index set of diagonal elements excluding the first entry equal to 1, which belongs already to ι_{g_1}. Second, we duplicate K times each u_l such that $u_{l,k}, k = 1, \ldots, K$, is the k-th duplicated element of u_l. Then, we can characterize the set ι_{g_3} of indices related to a scaled factor and its corresponding squared common instruments in the intercept as

$$\iota_{g_3} = \{\iota_{g_{3,1}}, \ldots, \iota_{g_{3,Kp}}\} \in \mathbb{N}_+^{Kp},$$

such that each set $\iota_{g_{3,j}} = \{u_{l,k}, d_1 + k + (l-1)\tilde{p} + 1\} \in \mathbb{N}^2_+$, $k = 1, \ldots, K$, can generate a single group containing two covariates and $\tilde{x}^{(3)}_{i,t} = (x_{i,t,j})_{j \in \iota_{g_3}} \in \mathbb{R}^{n_3}$, where $n_3 = 2Kp$. Finally, the last set ι_{g_4} of indices collects the indices related to Restrictions R.3 and R.4 for the stock-specific instruments $Z_{i,t-1}$ such that

$$\iota_{g_4} = \{\iota_{g_{4,1}}, \ldots, \iota_{g_{4,Kq}}\} \in \mathbb{N}_+^{Kq},$$

where each element $\iota_{g_{4,j}} = \{r_{m,k}, d_1+d_{21}+k+(m-1)q+1\} \in \mathbb{N}_+^{\tilde{p}+1}, m = 1, \ldots, q$, $k = 1, ..., K$, and $r_{m,k}$ is the k-th duplicated set of indices

$$r_{m,k} = \{d_{11} + m, \ldots, d_{11} + sq + m, \ldots, d_{11} + pq + m\} \in \mathbb{N}_+^{\tilde{p}+1},$$

for $s = 1, \ldots, \tilde{p}$, $k = 1, ..., K$. We define the last set of covariates groups as $\tilde{x}_{i,t}^{(4)} = (x_{i,t,j})_{j \in \iota_{g_4}} \in \mathbb{R}^{n_4}$, where $n_4 = Kq(\tilde{p} + 1)$. Next, we define the column vector

$$\tilde{x}_{i,t} = \left(\tilde{x}_{i,t}^{(1)\top}, \tilde{x}_{i,t}^{(2)\top}, \tilde{x}_{i,t}^{(3)\top}, \tilde{x}_{i,t}^{(4)\top}\right)^\top \in \mathbb{R}^{\tilde{d}},$$

where $\tilde{d} = \sum_{j=1}^4 n_j = K(\tilde{p}(q + 2) + q - 1) + (\tilde{p} - 1)\tilde{p}/2 + 1$. Let $\tilde{g} \in \widetilde{\mathcal{G}}$ denote a possible set of indices of the duplicated covariates $\tilde{x}_{i,t}$, where

$$\widetilde{\mathcal{G}} = \Big\{\iota_{g_1}, \iota_{g_{2,1}}, \ldots, \iota_{g_{2,n_2}}, \iota_{g_{3,1}}, \ldots, \iota_{g_{3,Kp}}, \iota_{g_{4,1}} \ldots, \iota_{g_{4,Kq}}\Big\}.$$

The sets $\mathcal{G}$ and $\widetilde{\mathcal{G}}$ are based on the original covariates $x_{i,t}$ for the former and the duplicated covariates $\tilde{x}_{i,t}$ for the latter, and we have that $J = |\mathcal{G}| = |\widetilde{\mathcal{G}}| = 1+n_2+Kp+Kq$.