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

Gaetan Bakalli and Stéphane Guerrier and Olivier Scaillet

January 2021

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.

G. Bakalli is with the Geneva School of Economics and Management, University of Geneva, Bd Pont-d'Arve 40, CH-1211 Geneva 4, Switzerland
(e-mail: gaetan.bakalli@unige.ch).
S. Guerrier is with the Faculty of Science & Geneva School of Economics and Management, University of Geneva, Bd Pont-d'Arve 40, CH-1211 Geneva 4, Switzerland.
(e-mail: stephane.guerrier@unige.ch).
O. Scaillet is with the Geneva Finance Research Institute, University of Geneva and Swiss Finance Institute, Bd Pont-d'Arve 40, CH-1211 Geneva 4, Switzerland (e-mail: 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. (2020) 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. 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 Group-LASSO with Overlap (OGL) of Jacob et al. (2009), 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, our OGL 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 studied within our model selection procedure. We derive an upper bound, and show that the number of explored models without grouping is divided by $2^3 = 8$, 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 explored without grouping. We exemplify this reduction with a simple two-factor example in Section 3.1. Consequently, the OGL 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 OGL 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) use penalisation techniques for prediction purposes. 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 penalisation 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 OGL 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 estimated support 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 performance 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 and the proofs of our theoretical results in Appendix B.

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 $1,\ldots,T$, 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 $\gamma_i \in [0,1], i = 1,\ldots,n$, with $\Sigma_{\varepsilon_t,t,n} \in \mathbb{R}^{n\times n}$ being the conditional variance-covariance matrix of the vector $(\varepsilon_t(\gamma_1),\ldots,\varepsilon_t(\gamma_n))^\top$ knowing Z_{t-1}, there exists a set such that $n^{-1}\,\mathrm{eig}_{\max}(\Sigma_{\varepsilon_t,t,n}) \xrightarrow{L^2} 0$ as $n \to \infty$, where $\mathrm{eig}_{\max}(\Sigma_{\varepsilon_t,t,n})$ 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 $Z_{\underline{t-1}}$ and $Z_{t-1}(\gamma)$, 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 $Z_{\underline{t}} = \{Z_t, Z_{t-1}, ...\}$ 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 $\mathbb{E}[f_t|\mathcal{F}_{t-1}]$ 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) $\mathbb{E}[f_t|\mathcal{F}_{t-1}] = F_0 + F_1 Z_{t-1}$, 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 $B(\gamma), C(\gamma), \Lambda_1$ 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 $Z_{i,t-1}$ drive the dynamics of the factor loadings $b_{i,t}$ for assets 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

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

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 Z_t. The vector-half operator $\operatorname{vech}(\cdot)$ 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 $\operatorname{vech}(X_t)$ is related to the time-invariant coefficients $A_i^\top(\Lambda_0 - F_0)$, whereas the elements $2, \ldots, \tilde{p}$ are related to $A_i^\top(\Lambda_1 - F_1)Z_{t-1} + Z_{t-1}^\top B_i^\top(\Lambda_0 - F_0)$. 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

$$\begin{aligned}
\beta_{1,i} &= \left(\beta_{11,i}^\top, \beta_{12,i}^\top\right)^\top \in \mathbb{R}^{d_1},\\
\beta_{11,i} &= N_{\tilde{p}}\left[(\Lambda - F)^\top \otimes I_{\tilde{p}}\right]\operatorname{vec}(\breve{B}_i^\top) \in \mathbb{R}^{d_{11}},\\
\beta_{12,i} &= W_{\tilde{p},q}\left[(\Lambda - F)^\top \otimes I_q\right]\operatorname{vec}(C_i^\top) \in \mathbb{R}^{d_{12}},\\
N_{\tilde{p}} &= \frac{1}{2}D_{\tilde{p}}^+(W_{\tilde{p}} + I_{\tilde{p}^2}) \in \mathbb{R}^{[(\tilde{p}+1)\tilde{p}/2+\tilde{p}q]\times\tilde{p}^2},\\
\beta_{2,i} &= \left(\beta_{21,i}^\top, \beta_{22,i}^\top\right)^\top = \left(\operatorname{vec}(\breve{B}_i^\top)^\top, \operatorname{vec}(C_i^\top)^\top\right)^\top \in \mathbb{R}^{d_2},
\end{aligned} \tag{6}$$

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 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 OGL approach of Jacob et al. (2009). Then, the coefficients of the selected β_i are estimated (post-OGL) through an OLS time-series regression as in GOS. The second-pass regression relies on the Weighted Least-Square (WLS) estimator of GOS to estimate the vector ν, and takes the LASSO estimator of Tibshirani (1996) to select and estimate the matrix F of coefficients.

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 $\mathcal{P}(\{1, \dots, d\})$ denotes the power set of $\{1, \dots, d\}$. 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 A: *The time-invariant coefficients belong to a single group, where no amount of shrinkage is applied.*

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

RESTRICTION C: *For instrument $\tilde{Z}_{t-1,l}$, for $l = 1, \dots, \tilde{p}$, if all its corresponding $\tilde{Z}_{t-1,l} f_{t,k}$, for $k = 1, \dots, K$, 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 $m = 1, \dots, q$, if all its corresponding $Z_{i,t-1,m} f_{t,k}$ for $k = 1, \dots, K$, 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 D: *For instrument $\tilde{Z}_{t-1,l}$, for $l = 1, \dots, \tilde{p}$, if at least one of its corresponding $\tilde{Z}_{t-1,l} f_{t,k}$, for $k = 1, \dots, K$, 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 $m = 1, \dots, q$, if at least one of its corre-*

sponding $Z_{i,t-1,m} f_{t,k}$, 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 A 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 B 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 B reflects this point, and imposes no specific group-structure to those covariates which are penalized individually. Restrictions C and D are critical in the model construction. 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. We illustrate this point below on an simple example with two factors, a single common instrument, and a single characteristic.

From the set of restrictions listed above, 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 A to D, we consider the Group-LASSO of Yuan and Lin (2006) and define the following sets of indices. The first group related to Restriction A always includes all covariates corresponding to the time-invariant contribution. Hence, we define $\tilde{x}_{i,t}^{(1)} = (x_{i,t,j})_{j \in \iota_{g_1}} \in \mathbb{R}^{n_1}$, where $n_1 = K + 1$, and ι_{g_1} is a set of indices such that,

$$\iota_{g_1} = \{1, d_1 + 1, \ldots, d_1 + k\tilde{p} + 1, \ldots, d_1 + (K-1)\tilde{p} + 1\} \in \mathbb{N}_+^{K+1}, \tag{7}$$

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

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

where each individual scalar index of ι_{g_2} generates a single group containing only one element $x_{i,t,j}$, with $j \in \iota_{g_2}$. To implement Restrictions C and D in the Group-LASSO framework, we would need to apply the same construction as described in

(7), i.e., group the corresponding scaled factors $(Z_{t-1,l}f_{t,1}, \ldots, Z_{t-1,l}f_{t,K})$ with their corresponding squared element $Z^2_{t-1,l}$ in $x_{2,i,t}$ and $x_{1,i,t}$, respectively. However, this construction would constrain the set of possible models. Indeed, 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 would remove all groups if at least one of those groups is not selected. Here, the groups do not yield to a partition for $\mathcal{G}$, and therefore the Group-LASSO does not select necessarily the predefined groups (due to non-differentiability of the penalty term). 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 $v = (v_{g_1}^\top, \ldots, v_{g_J}^\top)^\top \in \mathcal{V} = \prod_{j=1}^J \mathcal{V}_{g_j}$, where $J = |\mathcal{G}|$, $|\cdot|$ denotes the cardinality of a set and g_j, $j = 1, \ldots, J$, denotes the j-th element of $\mathcal{G}$. Then, we define the following OGL estimator:

$$\hat{\beta}_i^{\text{OGL}} = \underset{\beta_i \in \mathbb{R}^d}{\operatorname{argmin}} \sum_t \left(I_{i,t}R_{i,t} - \beta_i^\top I_{i,t}x_{i,t}\right)^2 + \delta\Omega_{\cup}^{\mathcal{G}}(\beta_i), \tag{8}$$

with the penalty term $\Omega_{\cup}^{\mathcal{G}}(\beta_i)$ defined as

$$\Omega_{\cup}^{\mathcal{G}}(\beta_i) = \min_{v \in \mathcal{V}} \sum_{g \in \mathcal{G}} w_g \|v_g\|_2, \quad \text{s.t.} \quad \beta_i = \sum_{g \in \mathcal{G}} v_g, \tag{9}$$

where w_g denotes the predefined weight associated to group g such that $\min_{g \in \mathcal{G}} w_g \geq 0$, and $\delta \geq 0$ corresponds to the hyperparameter driving the amount of shrinkage. The penalty term in (9) 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 (8) is the duplication of covariates put forward in Jacob et al. (2009), that we adapt to our setting below.

Let us describe the group structure needed within a regular Group-LASSO by replicating our covariates to solve the original OGL problem and ensuring that Restrictions C and D 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

$$\iota_{g_3} = \{\iota_{g_{3,1}}, \ldots, \iota_{g_{3,Kp}}\} \in \mathbb{N}_+^{Kp}, \tag{10}$$

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 $\tilde{x}_{i,t}^{(3)} = (x_{i,t,j})_{j \in \iota_{g_3}} \in \mathbb{R}^{n_3}$,

where $n_3 = 2Kp$. Finally, the last set ι_{g_4} of indices collects the indices related to Restrictions C and D for the stock-specific instruments $Z_{i,t-1}$ such that

$$\iota_{g_4} = \{\iota_{g_{4,1}}, \ldots, \iota_{g_{4,Kq}}\} \in \mathbb{N}_+^{Kq}, \tag{11}$$

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

$$\widetilde{\mathcal{G}} = \left\{\iota_{g_1}, \iota_{g_{2,1}}, \ldots, \iota_{g_{2,n_2}}, \iota_{g_{3,1}}, \ldots, \iota_{g_{3,Kp}}, \iota_{g_{4,1}} \ldots, \iota_{g_{4,Kq}}\right\}. \tag{12}$$

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. Since we plug the duplicated variables in the groups of $\mathcal{G}$, we do not change the number of possible groups, and we have $J = |\mathcal{G}| = |\widetilde{\mathcal{G}}| = 1 + n_2 + Kp + Kq$. Based on $\tilde{g}$, we let $\tilde{v}_{\tilde{g}} \in \widetilde{\mathcal{V}}_{\tilde{g}} = \{x \in \mathbb{R}^{\tilde{d}} | \operatorname{supp}(x) = \tilde{g}\}$, for $\tilde{g} \in \widetilde{\mathcal{G}}$ as well as $\tilde{v} = (\tilde{v}_{\tilde{g}_1}^\top, \ldots, \tilde{v}_{\tilde{g}_J}^\top)^\top \in \widetilde{\mathcal{V}} = \prod_{j=1}^J \widetilde{\mathcal{V}}_{\tilde{g}_j}$. The OGL problem defined in (8) can now be solved through the following equivalent optimization program:

$$\begin{aligned} \hat{\beta}_i^{\text{OGL}} = \underset{\beta_i \in \mathbb{R}^d}{\operatorname{argmin}} & \left\{ \sum_t \left(I_{i,t} R_{i,t} - \beta_i^\top I_{i,t} \tilde{x}_{i,t}\right)^2 + \delta \sum_{g \in \widetilde{\mathcal{G}}} w_g \|\tilde{v}_g\|_2 \right\}, \\ \text{s.t.} \quad & \beta_i = \sum_{g \in \widetilde{\mathcal{G}}} \tilde{v}_g. \end{aligned} \tag{13}$$

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 w_g is equal to zero. Every subset of $\mathcal{G}$ can be associated to a model. 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. 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 LASSO method examines many more possibilities. Besides, from Assumption A.1, we have $\min(p, q) \geq 1$, and deduce the upper bound:

$$\frac{2^{J-1}}{2^{d-n_1}} \leq \frac{1}{8}. \tag{14}$$

To illustrate the grouping structure and the importance of Restrictions A to D, 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 $\widetilde{\mathcal{G}}$ in (12), 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 ι_{g_1}, $(x_{1,i,t,2})$ for $\iota_{g_{2,1}}$, $(x_{1,i,t,3}, x_{2,i,t,2})^\top$ for $\iota_{g_{3,1}}$, $(x_{1,i,t,3}, x_{2,i,t,4})^\top$ for $\iota_{g_{3,2}}$, $(x_{1,i,t,4}, x_{1,i,t,5}, x_{2,i,t,5})^\top$ for $\iota_{g_{4,1}}$, and finally $(x_{1,i,t,4}, x_{1,i,t,5}, x_{2,i,t,6})^\top$ for $\iota_{g_{4,2}}$. Stacking those vectors line-wise in a single column defines the full vector of covariates $\tilde{x}_{i,t}$ for the OGL 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 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 LASSO 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 explores the set $\mathcal{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_{32}\}$ of possible models that respect Restrictions A to D with $\mathcal{M}_1$ being the model with the time-invariant contribution only (Assumption A.3). The OGL method gives 2^5 possible models. It is considerably smaller than the $2^8 = 256$ possible models explored by the LASSO method. Here, we reach the upper bound (14) 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 OGL approach, strengthening conducive arguments for our proposed method.

Let us define the true support of β_i as $\mathcal{S}_i = \text{supp}(\beta_i) \subseteq \{1, \ldots, d\}$. The goal of (13) is to recover the true support $\mathcal{S}_i$ of β_i, which is discussed in Jacob et al. (2009) under the following assumptions.

ASSUMPTION A.4: *(Group-support recovery Jacob et al. (2009))*
(i) $^1/_{T_i} \sum_t I_{i,t} x_{i,t} x_{i,t}^\top$ *is a positive definite matrix.*
(ii) There exists a neighborhood of β_i *for which (9) has a unique solution.*

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

Table 1: Set of possible models according to Restrictions A-D 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 $x_{l,i,t,k} = x_{l,k}$

Assumption A.4 i) is a standard regularity condition. Assumption A.4 ii) requires the true support for asset i to be unique and is discussed in Jacob et al. (2009). Under Assumptions A.4, $\delta_{T_i} \to 0$ with $\delta_{T_i} T_i^{1/2} \to \infty$, Conditions C1 and C2 of Jacob et al. (2009), $\mathcal{S}_i$ is asymptotically contained in $\hat{\mathcal{S}}_i$, i.e.,

$$\Pr(\mathcal{S}_i \subseteq \hat{\mathcal{S}}_i) \to 1, \tag{15}$$

where $\hat{\mathcal{S}}_i$ is the estimated support of the estimated coefficients for asset i. Conditions C1 and C2 are discussed in Jacob et al. (2009) and required for β_i to be a feasible solution of (8). Let us further define the set of estimated supports for all $i = 1, \ldots, n$, as $\hat{\mathcal{S}} = (\hat{\mathcal{S}}_1, \ldots, \hat{\mathcal{S}}_n)$ and its set of true values $\mathcal{S} = (\mathcal{S}_1, \ldots, \mathcal{S}_n)$. Those definitions are useful to derive the asymptotic properties.

In our approach, the OGL estimator recovers the support of β_i, for all i. Then, to estimate the vector parameter β_i in (5), we follow Feng et al. (2020), and rely on a post-OGL estimator, regressing the $x_{i,t}$ included in the estimated set for each asset i. Post-LASSO approaches are now standard (see the review paper of Chernozhukov et al. (2015)). For that purpose, we introduce the indicator vector $\mathbf{1}_{\beta_i} \in \mathbb{N}^d$, such that $\mathbf{1}_{\beta_{i,j}} = 1$ if $\beta_{i,j} \neq 0$, and 0 otherwise, for $j = 1, \ldots, d$, 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 OGL approach. Hence, in the same spirit as in Chaieb et al. (2020), 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 $\text{diag}(\mathbf{1}_{\beta_{22,i}})$. We can now introduce the post-OGL covariates and parameter specification of dimension $d_{1,i} = d_{1,i} + d_{2,i}$, where $d_{1,i} = d_{11,i} + d_{12,i}$ and $d_{2,i} = d_{21,i} + d_{22,i}$ such that $\check{x}_{i,t} = \left(\check{x}_{1,i,t}^\top, \check{x}_{2,i,t}^\top\right)^\top$, where

$$\check{x}_{1,i,t} = \left(\text{vech}\left[X_t\right]^\top \tilde{D}_i, \left(\tilde{Z}_{t-1}^\top \otimes Z_{i,t-1}^\top\right) \tilde{E}_i\right)^\top \in \mathbb{R}^{d_{1,i}},$$

$$\check{x}_{2,i,t} = \left(\left(\tilde{Z}_{t-1}^\top \otimes f_t^\top\right) \tilde{B}_i^\top, \left(Z_{i,t-1}^\top \otimes f_t^\top\right) \tilde{C}_i^\top\right)^\top \in \mathbb{R}^{d_{2,i}}.$$

Based on this definition of $\check{x}_{i,t}$, we can finally define the post-OGL vector parameter $\check{\beta}_i = (\check{\beta}_{1,i}^\top, \check{\beta}_{2,i}^\top)^\top$, where

$$\check{\beta}_{1,i} = \Bigg(\tilde{D}_i^\top N_{\tilde{p}} \left[(\Lambda - F)^\top \otimes I_{\tilde{p}}\right] \tilde{B}_i^\top \tilde{B}_i \,\text{vec}\left[\check{B}_i^\top\right], \\ \tilde{E}_i^\top W_{\tilde{p},q} \left[(\Lambda - F)^\top \otimes I_q\right] \tilde{C}_i^\top \tilde{C}_i \,\text{vec}\left[C_i^\top\right]\Bigg)^\top,$$

$$\check{\beta}_{2,i} = \left(\tilde{B}_i \,\text{vec}\left[\check{B}_i^\top\right]^\top, \tilde{C}_i \,\text{vec}\left[C_i^\top\right]^\top\right)^\top,$$

yielding the linear regression model defined in terms of the sparse regressors $\check{x}_{i,t}$:

$$R_{i,t} = \check{\beta}_i^\top \check{x}_{i,t} + \varepsilon_{i,t}.$$

We can implement our post-OGL estimator on the updated first-pass regression, on the selected support $\hat{\mathcal{S}}_i$ and define the estimator of $\check{\beta}_i$ as $\hat{\beta}_i(\hat{\mathcal{S}}_i) = \hat{Q}_{\check{x},i}^{-1} \frac{1}{T_i} \sum_t I_{i,t} \check{x}_{i,t} R_{i,t}$, $i = 1, ..., n,$, where $\hat{Q}_{\check{x},i} = \frac{1}{T_i} \sum_t I_{i,t} \check{x}_{i,t} \check{x}_{i,t}^\top$. To control for short sample size, and potentially numerical instability on the inversion of matrix $\hat{Q}_{\check{x},i}$, we consider the trimming device defined in GOS, such that $\mathbf{1}_i^\chi = \mathbf{1}\{CN(\hat{Q}_{\check{x},i}) \leq \chi_{1,T}, \tau_{i,T} \leq \chi_{2,T}\}$, where $CN(\hat{Q}_{\check{x},i}) = \sqrt{\text{eig}_{\max}(\hat{Q}_{\check{x},i}) / \text{eig}_{\min}(\hat{Q}_{\check{x},i})}$ is the condition number of the matrix $\hat{Q}_{\check{x},i}$, $\text{eig}_{\min}(\cdot)$ denotes the minimum eigenvalue, and $\tau_{i,T} = T/T_i$. The first trimming based on $CN(\hat{Q}_{\check{x},i}) \leq \chi_{1,T}$ 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 $i = 1, \ldots, n$. Based on the selection matrices $\tilde{D}_i, \tilde{E}_i, \tilde{B}_i$, and $\tilde{C}_i$, we re-write the parameter restriction in (2) such that

$$\check{\beta}_{3,i} = \left(\left[\tilde{D}_i^\top N_{\tilde{p}} \left(\check{B}_i^\top \otimes I_{\tilde{p}} \right) \right]^\top, \left[\tilde{E}_i^\top W_{\tilde{p},q} \left(C_i^\top \otimes I_p \right) \right]^\top \right)^\top,$$

where $N_{\tilde{p}}$ is defined in (6), yielding the asset pricing restrictions expressed in the newly defined $\check{\beta}_{1,i}$ and $\check{\beta}_{3,i}$ as $\check{\beta}_{1,i} = \check{\beta}_{3,i}\nu$, $\nu = \text{vec}(\Lambda^\top - F^\top)$. We obtain $\check{\beta}_{3,i}$ from the following identity,

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

We can now implement the following second pass regression WLS estimator

$$\hat{\nu}(\hat{\mathcal{S}}) = \hat{Q}_{\beta_3}^{-1} \frac{1}{n} \sum_i \hat{\beta}_{3,i}(\hat{\mathcal{S}}_i)^\top \hat{w}_i \hat{\beta}_{1,i}(\hat{\mathcal{S}}_i), \tag{16}$$

where $\hat{\nu}(\hat{\mathcal{S}})$ denotes the estimator of $\hat{\nu}$ under the set of estimated support $\hat{\mathcal{S}}$, $\hat{Q}_{\beta_3} = \frac{1}{n} \sum_i \hat{\beta}_{3,i}(\hat{\mathcal{S}}_i)^\top \hat{w}_i \hat{\beta}_{3,i}(\hat{\mathcal{S}}_i)$, and weights are estimates of $w_i = \mathbf{1}_i^\chi \left(\text{diag}\left[v_i \right] \right)^{-1}$. For simplicity of notation we define the estimator $\hat{\beta}_i(\hat{\mathcal{S}}_i) = \hat{\beta}_i$, $\hat{\beta}_{1,i}(\hat{\mathcal{S}}_i) = \hat{\beta}_{1,i}$ and $\hat{\beta}_{3,i}(\hat{\mathcal{S}}_i) = \hat{\beta}_{3,i}$. 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 $v_i = \tau_i C_{\nu,1,i}^\top Q_{\check{x},i}^{-1} S_{ii} Q_{\check{x},i}^{-1} C_{\nu,1,i}$ where $Q_{\check{x},i} = \mathbb{E}[\check{x}_{i,t} \check{x}_{i,t}^\top | \gamma_i]$. Moreover, we have that $S_{ii} = \text{plim}_{T\to\infty} \frac{1}{T} \sum_t \sigma_{ii,t} \check{x}_{i,t} \check{x}_{i,t}^\top = \mathbb{E}[\varepsilon_{i,t}^2 \check{x}_{i,t} \check{x}_{i,t}^\top | \gamma_i]$ with $\sigma_{ii,t} = \mathbb{E}[\varepsilon_{i,t}^2 | \check{x}_{i,\underline{t}}, \gamma_i]$ 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 $\hat{v}_i = \tau_{i,T} C_{\hat{\nu}_1}^\top \hat{Q}_{\check{x},i}^{-1} \hat{S}_{ii} \hat{Q}_{\check{x},i}^{-1} C_{\hat{\nu}_1}$, where $\hat{S}_{ii} = \frac{1}{T_i} \sum_t I_{i,t} \hat{\varepsilon}_{i,t}^2 \check{x}_{i,t} \check{x}_{i,t}^\top$, $\hat{\varepsilon}_{i,t} = R_{i,t} - \hat{\beta}_i^\top \check{x}_{i,t}$ and $C_{\hat{\nu},1,i} = \left(E'_{1,i} - \left(I_{d_{1,i}} \otimes \hat{\nu}_{1,i}^\top\right) J_{a,i} E_{2,i}^\top\right)^\top$. To estimate $C_{\nu,1,i}$, we use the OLS estimator $\hat{\nu}_{1,i} = \left(\sum_i \mathbf{1}_i^\chi \hat{\beta}_{3,i}^\top \hat{\beta}_{3,i}\right)^{-1} \sum_i \mathbf{1}_i^\chi \hat{\beta}_{3,i}^\top \hat{\beta}_{1,i}$. We estimates the weights through $\hat{w}_i = \mathbf{1}_i^\chi \left(\text{diag}\left[\hat{v}_i\right]\right)^{-1}$.

To study the asymptotic properties of the estimator $\hat{\nu}(\hat{\mathcal{S}})$, we consider the following assumptions on the dependence structure and size of the cross-section n.

ASSUMPTION A.5: *(Conditional heteroskedasticity)*
There exists a positive constant M such that for all n, T, $\frac{1}{M} \leq \sigma_{ii,t} \leq M$, $i = 1, ..., n$.

ASSUMPTION A.6: *(Relative rates and bounds)*
i) The size of the cross section is such that $n = \mathcal{O}(T^{\bar{\gamma}})$ for $\bar{\gamma} > 0$.
ii) The probability that the estimated set $\hat{\mathcal{S}}_i$ is not included in the true set of active group $\mathcal{S}_i$ is such that $\Pr(\mathcal{S}_i \not\subseteq \hat{\mathcal{S}}_i) = \mathcal{O}(T^{-\omega})$ for $\omega > 0, \omega > \bar{\gamma}$ and $i = 1, \ldots, n$.

Assumption A.5 allows for potential conditional heteroskedasticity in the error terms. Assumption A.6 i) 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, while Assumption A.6 b) requires the probability that the true support $\mathcal{S}_i$ is not included in the estimated support $\hat{\mathcal{S}}_i$, for all $i = 1, \ldots, n$, converges to 0 at a faster rate than n diverges. In Proposition 1, we provide the consistency result for the estimator $\hat{\nu}(\hat{\mathcal{S}})$.

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

This asymptotic property of $\hat{\nu}(\hat{\mathcal{S}})$ is studied under the double asymptotics $n, T \to \infty$. Lemma 2 in Appendix B shows that it yields $\Pr(\mathcal{S} \subseteq \hat{\mathcal{S}}) \to 1$. The estimator $\hat{\nu}(\hat{\mathcal{S}})$ is therefore consistent for ν under the estimated support $\hat{\mathcal{S}}$. GOS 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. Proof of Proposition 1 is given in Appendix B.

Let us now recover the sparse structure of the conditional expectation of the factors under Assumption A.2. For that purpose, we consider the LASSO estimator of Tibshirani (1996) 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:

$$\hat{F}_k = \underset{F_k \in \mathbb{R}^{\tilde{p}}}{\text{argmin}} \sum_t \left(f_{k,t} - F_k^\top \tilde{Z}_{t-1}\right)^2 + \delta \|F_k\|_2, \tag{17}$$

where δ accounts for the amount of shrinkage. The estimate $\hat{F}$ stacks row-wise the elements of $\hat{F}_k$ obtained from (17). 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 $\text{vec}(\hat{\Lambda}^\top) = \hat{\nu}(\hat{\mathcal{S}}) + \text{vec}(\hat{F}^\top)$, 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 $\hat{\nu}(\hat{\mathcal{S}})$ and the work of Knight and Fu (2000), which study the asymptotic properties of the LASSO estimator under the following assumptions:

ASSUMPTION A.7: *(Consistency of LASSO Knight and Fu (2000))*
i) $W_T = {}^1\!/_T \sum_{t=1}^{T} \tilde{Z}_{t-1}\tilde{Z}_{t-1}^\top \xrightarrow{\mathcal{P}} W$*, where* W *is a positive definite matrix.*
ii) W_T *is a non-singular matrix.*

Assumptions A.7 are standard regularity assumptions on the design matrix for linear regression model, in order to obtain a unique solution for F_k. Under the above Assumption A.7, Knight and Fu (2000) show that, for a sequence of δ_T such that ${}^{\delta_T}\!/_T \to \delta_0$ where $\delta_0 \geq 0$, $\hat{F}_k \xrightarrow{\mathcal{P}} F_k$ as $T \to \infty$, and with F_k being the true value of the vector parameter $\hat{F}_k$. Based on this results and Proposition 1, the following Proposition gives the consistency result for the estimator $\hat{\Lambda}$.

PROPOSITION 2: *(Consistency of* $\hat{\Lambda}$*)*
From Proposition 1, under APR.1 to APR.4, SC.1 and SC.2 of GOS, Assumptions A.2, A.7 and 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 $\hat{\Lambda}$, $\|\operatorname{vec}(\hat{\Lambda}^\top - \Lambda^\top)\| \leq \|\hat{\nu}(\hat{\mathcal{S}}) - \nu\| + \|\operatorname{vec}(\hat{F}^\top - F^\top)\|$. From Proposition 1, we know that $\|\hat{\nu}(\hat{\mathcal{S}}) - \nu\| = o_p(1)$ and from Assumptions A.7 and B.6 $\|\operatorname{vec}(\hat{F}^\top - F^\top)\| = o_p(1)$. Hence consistency of $\hat{\lambda}_t$, $\sup_t \|\hat{\lambda}_t - \lambda_t\| = o_p(1)$, is implied under Assumption 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 OGL method and at comparing it with the LASSO 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 $\sigma^2 = 0.15$. 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. This setting corresponds to a theoretical R^2 of 51% and a signal-to-noise ratio of approximately 1.05.

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). The resulting β_i has 24 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 OGL estimator.

The selection and prediction performance is measured through the median of the Mean Prediction Error (Med(PE_R)), the median of Mean Squared Error (MSE) for parameter β_i (Med(MSE_β)), the proportion of times the model introduces arbitrage (Arb. (%)), the proportion of times the correct model is nested within the selected model (Inc. (%)), 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 post-OGL method makes a better job at predicting out-of-sample with a reduction of 22% w.r.t. the post-LASSO method. The improvement in the median of the MSE for β_i is 75%. Contrary to the LASSO methods, for which 96.8% of estimated models exhibit arbitrage, the OGL methods select only models without introducing *ex-ante* arbitrage by construction. Since we face less than 100% for the LASSO methods, they sometimes shrink adequately to zero the coefficients that should be. Contrary to the LASSO methods, the OGL methods are almost always (96.2%) included in the true model as expected from our theory, which predicts 100% asymptotically. The LASSO methods yield a zero proportion since as soon as it does not shrink to zero a coefficient that should be because of the no-arbitrage restriction, the LASSO methods fail. The OGL methods are able to recover the 24 true non-zero coefficients (23.7) while the LASSO methods struggle (11.5). They are also more parsimonious in terms of selected regressors (52.8 versus 74.5). Overall, the performance of the OGL methods is much closer to the oracle where we estimate the true DGP with the known sparsity by OLS (oracle-OLS).

Our second simulation set-up focuses on the out-of-sample prediction performance of the post-OGL method in a setting close to our empirical study. 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 post-OGL 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 $\text{corr}(\varepsilon_{k,t}, \varepsilon_{l,t}) = 0.25^{|k-l|}$, $k, l = 1, ..., 50, l \neq k$. The variance of each error $\varepsilon_{i,t}$ is set equal to 0.05. From those 500 simulated paths, we implement the OGL estimation procedure of Section 3, and compare it with the same procedure, but using the LASSO estimator instead of the OGL estimator to select the covariates in (5). To evaluate the

Method	Med (PE_R)	Med(MSE_β)	Arb. (%)	Inc. (%)	true +	NbReg
OGL	$3.3 \cdot 10^{-2}$	$1.1 \cdot 10^{-1}$	0.0	96.2	23.7	52.8
post-OGL	$3.0 \cdot 10^{-2}$	$1.3 \cdot 10^{-1}$	0.0	96.2	23.7	52.8
LASSO	$3.1 \cdot 10^{-2}$	$3.5 \cdot 10^{-1}$	98.8	0.0	11.5	74.5
post-LASSO	$3.8 \cdot 10^{-2}$	$3.7 \cdot 10^{-1}$	98.8	0.0	11.5	74.5
oracle-OLS	$2.4 \cdot 10^{-2}$	$4.8 \cdot 10^{-2}$	0.0	100.0	24.0	24.0

Table 2: Performance of estimation and model selection criteria. The methods include the OGL, post-OGL, LASSO, post-LASSO, and oracle-OLS. We simulate 500 samples under the true sparse DGP. We report the median of the Mean Prediction Error (Med(PE_R)), the median of Mean Squared Error (MSE) for parameter β_i (Med(MSE_β)), the proportion of times the model does not introduce arbitrage (Arb. (%)), the proportion of times the correct model is nested within the selected model (Inc. (%)), the average number of selected true non-zero coefficients (true +), and average number of regressors in the selected model (NbReg).

Methods	RMSPE	Av(|PE|)	Std(PE)
post-OGL	$1.0 \cdot 10^{-2}$	$9.9 \cdot 10^{-3}$	$3.1 \cdot 10^{-4}$
post-LASSO	$1.3 \cdot 10^{-2}$	$1.3 \cdot 10^{-2}$	$4.5 \cdot 10^{-4}$

Table 3: Out-of-sample prediction performance of an equally-weighted portfolio. We compare the post-OGL and post-LASSO methods. We simulate excess return paths for 500 assets under sparse DGPs. We report Root Mean Squared Prediction Error (RMSPE), Mean Absolute Prediction Error (Av(|PE|)) and Standard Deviation of the Prediction Error (Std(PE)) of an equally-weighted portfolio.

out-of-sample prediction performance, we simulate one new cross-sectional sample (testing sample) from the 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 post-OGL and post-LASSO 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), the Mean Absolute Prediction Error (Av(|PE|)), and the Standard Deviation (Std(PE)) of the Prediction Error over the vector gathering the PE at each out-of-sample date. We repeat this procedure 50 times to get an average. They are reported in Table 3. We can see that the post-OGL method is much better at out-of-sample predicting excess returns of an equally-weighted portfolio both in terms of average $|PE|$ (reduction by 24%) but also in terms of variability as measured by Std(PE) (reduction by 32%). The empirical distribution of the prediction errors is given in Figure 1. We can see that the post-OGL method is centered on zero with a lower variance contrary to the post-LASSO method. Those second simulation results again point in favor of our advocated selection method.

5 Empirical results

This section investigates the predictive capacity of the post-OGL estimator and compares it with the post-LASSO estimator, and a pure time-invariant model. We use the post-LASSO estimator to gauge the added value of incorporating the no-arbitrage restrictions in the penalisation approach and the time-invariant model to gauge the added value of allowing for time-variation.

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. (2020)). 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

In this section, we compare the in-sample prediction performance of the penalized two-pass procedure with OGL described in Section 3 to the two following methods. The first method computes the estimator of GOS in the first and second-pass regressions for a *time-invariant* model (time-invariant contribution only, see Assumption A.3) with only the factors f_t as regressors. Our second estimator is the post-LASSO estimator, where we select the $x_{i,t}$ in the first-pass regression with the LASSO estimator of Tibshirani (1996) and fit the WLS estimator for the ν described in Section 2, while computing the estimator $\hat{F}$ as in (17). 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 searched by the LASSO method is 2^{194} (2^{213}) with $K = 4$ ($K = 5$), while the number of models searched by the OGL method is 2^{97} (2^{116}), which gives the ratio 2^{-97}, a much lower value than the upper bound $1/8$ in (14).

We choose the regularisation parameter δ in a data dependent way to minimize the Akaike Information Criterion (AIC) for both post-OGL and post-LASSO 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 $\mathbf{1}\{CN(\hat{Q}_{\tilde{x},i}) \leq \chi_{1,T}\}$, 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 post-OGL estimator, the post-LASSO estimator, and the time-invariant estimator, we end up with 4582, 4238, 4879 for the four-factor model, and 4549, 4176, 4879 for the five-factor model. We can observe that the trimming device for the post-LASSO method is more binding, since as seen in the simulation results in Section 4, the post-LASSO method tends to includes more variables, and increases 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 A related to Assumption A.3. More than half of the stocks require dynamics in their factor loadings. This new empirical result based on a penalisation 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 (Arbitrage (%)) of estimated models with time-varying loadings and presenting arbitrage, namely selecting covariates violating the no-arbitrage restrictions. For that computation, we exclude the time-invariant estimates and OGL estimates, which both avoid *ex-ante* arbitrage by construction. In line with our Monte Carlo results, the post-LASSO procedure ends up with all the time-varying models estimated with arbitrage for both specifications. We conclude that the post-OGL estimation achieves parsimony while avoiding arbitrage in time-varying factor models.

	Carhart four-factor		Fama-French five-factor	
Methods	TI (%)	Arbitrage (%)	TI (%)	Arbitrage (%)
post-OGL	42	0	49	0
post-LASSO	46	100	44	100
time-invariant	100	0	100	0

Table 4: Percentage (TI (%)) of estimated models shrunk towards the time-invariant specification and percentage (Arbitrage (%)) of estimated time-varying models presenting arbitrage with the Carhart four-factor and Fama-French five-factor models for the post-OGL, post-LASSO, and time-invariant methods. The sample of US equity excess returns begins in July 1963 and ends in December 2019.

Let us now investigate in-sample predictability performance. As, in Chaieb et al. (2020), we decompose the conditional expected return of asset i for month t for both time-varying factor specifications, as:

$$\mathbb{E}\left[R_{i,t}|\mathcal{F}_{t-1}\right] = a_{i,t} - b_{i,t}^{\top}\nu_t + b_{i,t}^{\top}\lambda_t = a_{i,t} + b_{i,t}^{\top}\mathbb{E}[f_t|\mathcal{F}_{t-1}]. \tag{18}$$

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. (2020) show that incorporating pricing errors, instead of only relying on $b_{i,t}^{\top}\lambda_t$ in (18), helps to predict future equity excess returns. Similarly, for the time-invariant models, we decompose the unconditional expected return as:

$$\mathbb{E}\left[R_{i,t}\right] = a_i - b_i^{\top}\nu + b_i^{\top}\lambda = a_i + b_i^{\top}\mathbb{E}[f_t]. \tag{19}$$

For such time-invariant specifications, the contribution of the pricing errors $a_i - b_i^{\top}\nu$ is often large. To compare the prediction performance of the three 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. (2020) also uses this weighting scheme. For that portfolio, we compute the PE by comparing the prediction made at time t by each model ((18) and (19)) to the forward 12-months realized excess returns, namely the average of the realized excess returns over the next 12 months. Table 5 reports the RMSPE, as well as the Av(|PE|) and Std(PE) of the Prediction Error for the Carhart four-factor model and Fama-French five-factor model specifications. The post-OGL method performs better than its competitors even for that very diversified stable portfolio, where we expect differences in prediction performance to be attenuated. In particular, the Av(|PE|) is reduced by 10% to 20%. Figure 2 displays the corresponding box-plots of the PE computed at each month for each method. The box-plots for the post-OGL method in Figure 2 are narrower when compared to the two other methods as the PE are more concentrated around zero. Those predictability improvements provide further evidence in support for the advocated post-OGL approach in the first-pass regression on top of the need to incorporate model parameter restrictions to get models compatible *ex-ante* with the no-arbitrage restrictions.

	Carhart four-factor			Fama-French five-factor		
Methods	RMSPE	Av(|PE|)	Std(PE)	RMSPE	Av(|PE|)	Std(PE)
post-OGL	$1.4 \cdot 10^{-2}$	$1.1 \cdot 10^{-2}$	$3.1 \cdot 10^{-4}$	$1.4 \cdot 10^{-2}$	$1.1 \cdot 10^{-2}$	$3.2 \cdot 10^{-4}$
post-LASSO	$1.5 \cdot 10^{-2}$	$1.2 \cdot 10^{-2}$	$3.1 \cdot 10^{-4}$	$1.6 \cdot 10^{-2}$	$1.3 \cdot 10^{-2}$	$3.3 \cdot 10^{-4}$
time-invariant	$1.5 \cdot 10^{-2}$	$1.2 \cdot 10^{-2}$	$3.3 \cdot 10^{-4}$	$1.5 \cdot 10^{-2}$	$1.4 \cdot 10^{-2}$	$3.3 \cdot 10^{-4}$

Table 5: Root Mean Squared Prediction Error (RMSPE), Mean Absolute Prediction Error (Av(|PE|)) and Standard Deviation of the Prediction Error (Std(PE)) of an equally-weighted portfolio with the Carhart four-factor and Fama-French five-factor models for the post-OGL, post-LASSO, and time-invariant methods. The sample of US equity excess returns begins in July 1963 and ends in December 2019.

To further investigate time-varying predictability, Figures 4 and 5 show the forward 12-months realized excess returns for the equally-weighted portfolio and compare them with the predicted excess returns computed from (18) and (19) for the two methods with penalisation. In both Figures 4 and 5, the post-OGL and post-LASSO predicted excess return paths (red plain line) overall reconcile well with the realized excess returns (black dashed line). However, the post-LASSO method 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 LASSO 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 post-OGL 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 post-OGL method as it should be if we believe that the factors are most of the time fully tradable. Figures for the Carhart four-factor model are similar, and thus omitted.

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 than in 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. We only report the results for the five-factor model of Fama and French (2015), since the results for the four-factor model of Carhart (1997) are similar.

	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)
post-OGL	$1.6 \cdot 10^{-2}$	$1.3 \cdot 10^{-3}$	$3.5 \cdot 10^{-4}$	$1.0 \cdot 10^{-2}$	$8.2 \cdot 10^{-3}$	$2.0 \cdot 10^{-4}$
post-LASSO	$1.6 \cdot 10^{-2}$	$1.2 \cdot 10^{-3}$	$3.9 \cdot 10^{-4}$	$1.4 \cdot 10^{-2}$	$1.2 \cdot 10^{-2}$	$1.8 \cdot 10^{-4}$
time-invariant	$1.7 \cdot 10^{-2}$	$1.4 \cdot 10^{-3}$	$3.9 \cdot 10^{-4}$	$1.6 \cdot 10^{-2}$	$1.2 \cdot 10^{-2}$	$3.8 \cdot 10^{-4}$

Table 6: Out-of-sample Root Mean Squared Prediction Error (RMSPE), Mean Absolute Prediction Error (Av(|PE|)) and Standard Deviation of the Prediction Error (Std(PE)) of an equally-weighted portfolio with the Fama-French five-factor model specification for the post-OGL, post-LASSO, and time-invariant 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.

Table 6 reports the out-of-sample results for an equally-weighted portfolio and the two testing periods. Figures 6 and 7 show the forward 12-months realized excess returns for the equally-weighted portfolio and compare them with the predicted excess returns out-of-sample. Overall, we confirm the messages conveyed by the in-sample analysis, and we do not repeat all of them to save space. The post-OGL method beats the two other methods in out-of-sample prediction performance on the recent testing period 2010-2019. In Table 6, the (Av(|PE|)) is reduced by 32% when compared to the two other methods. 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 also have a reduction of 8% for the older testing period 2000-2009 when compared to the time-invariant method. The Post-LASSO method does better by 8% in average |PE|, but its RMSPE is equal to the one of the post-OGL method. The older testing period includes two recession periods with large swings in the realized portfolio excess returns. As a consequence, the predictability performance deteriorates. For both testing periods, the box-plots in Figure 3 show that out-of-sample PE related to the portfolio excess returns for the post-OGL method are located closer to zero and more symmetrically distributed. Their scale is narrower for the older period and comparable for the recent period. As observed in the in-sample analysis, the post-OGL method seems to perform better in terms of out-of-sample predictability.

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 are grateful to I. Chaieb, H. Langlois, A.-P. Fortin, and P. Zaffaroni for their helpful comments, as well as participants at the 8th days of Econometrics for Finance, 2021 SoFiE Machine Learning conference, and Vienna Workshop "Econometrics of Option Markets". G. Bakalli and O. Scaillet were supported by the SNSF Grant $\#100018-182582$. S. Guerrier was supported in part by the SNSF Professorships Grant #176843 and by the Innosuisse-Boomerang Grant #37308.1 IP-ENG.

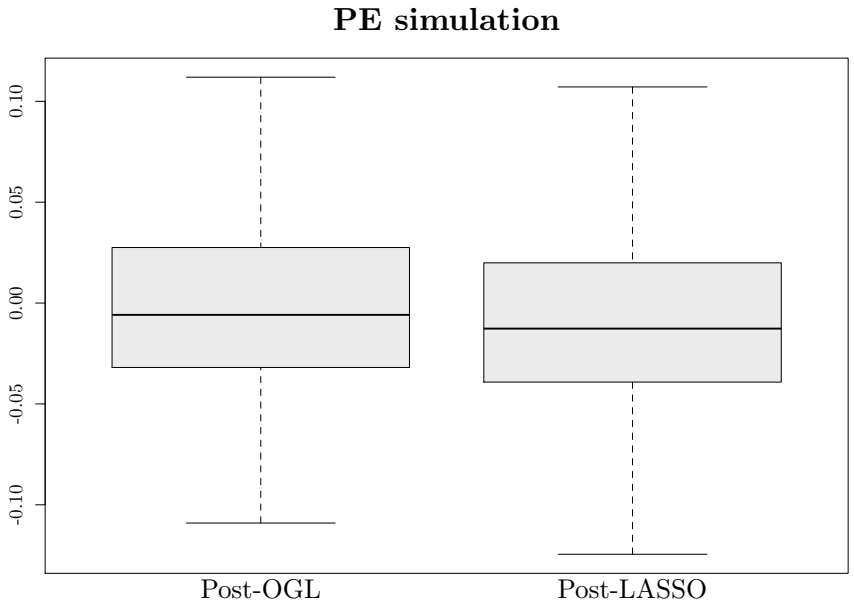

Figure 1: Empirical distribution of out-of-sample Prediction Error (PE) of an equally-weighted portfolio. We compare the post-OGL and post-LASSO methods. We simulate excess return paths for 500 assets under sparse DGPs.

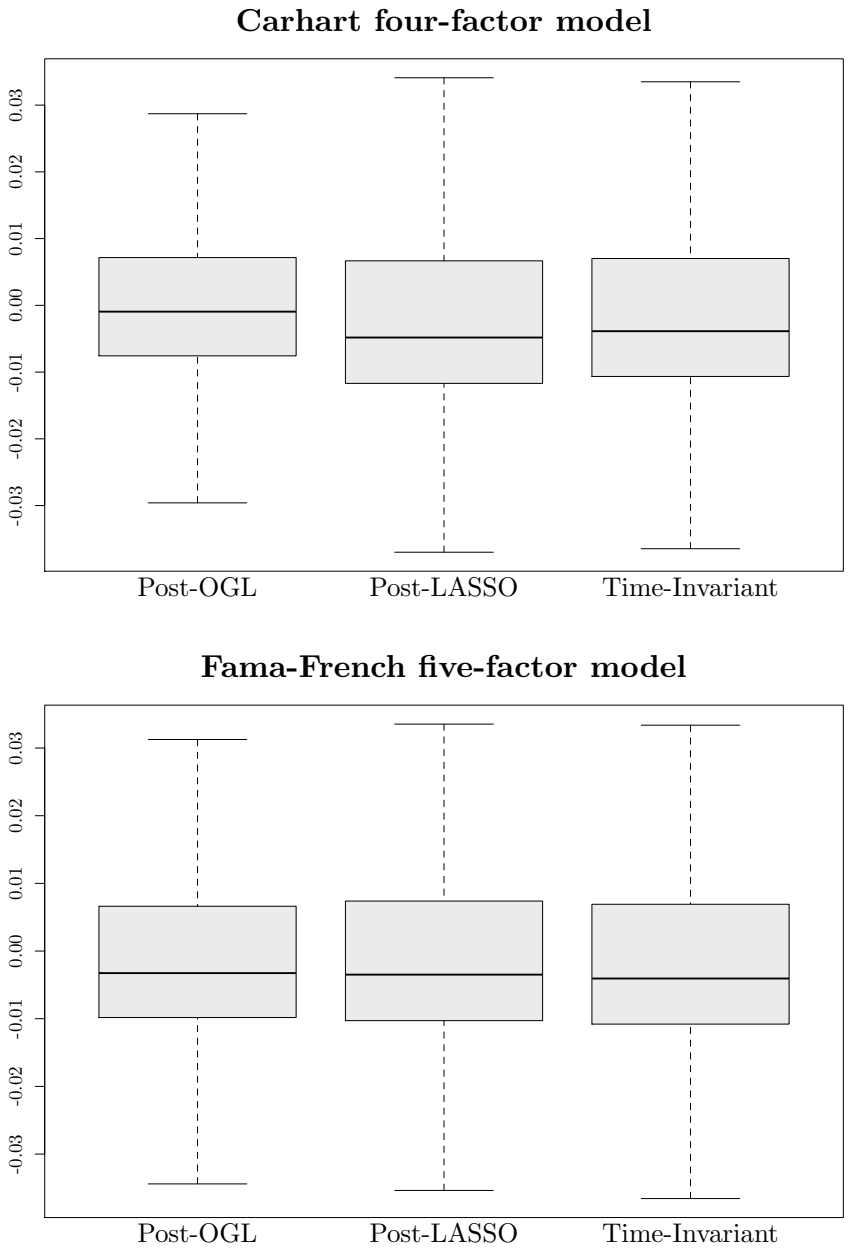

Figure 2: Empirical distribution of in-sample Prediction Error (PE) of an equally-weighted portfolio. We compare the post-OGL, post-LASSO, and time-invariant methods. The upper panel corresponds to the Carhart four-factor model. The lower panel corresponds to the Fama-French five-factor model. The sample of US equity excess returns begins in July 1963 and ends in December 2019.

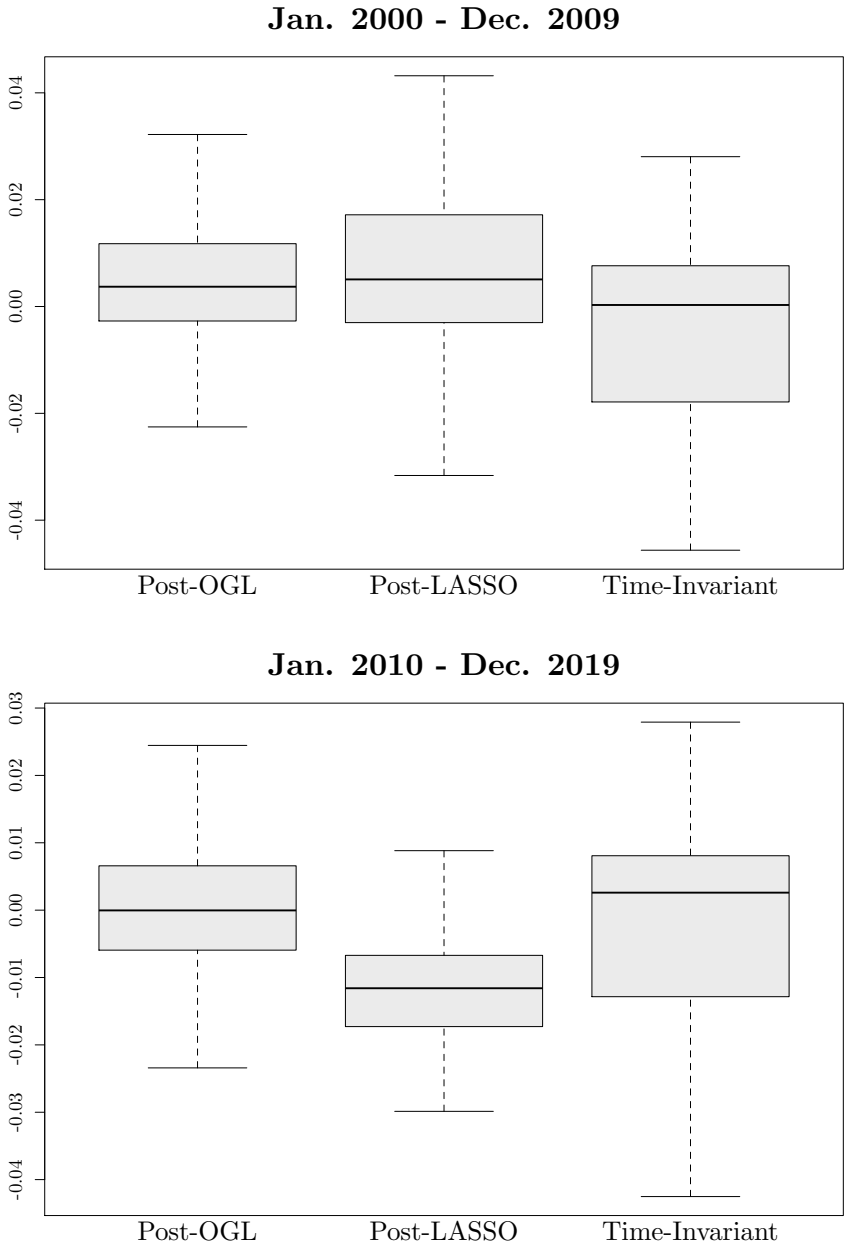

Figure 3: Empirical distribution of out-of-sample Prediction Error (PE) of an equally-weighted portfolio. We compare the post-OGL, post-LASSO, and time-invariant methods for the Fama-French five-factor model. The upper panel is for the testing period 2000-2009. The lower panel is for the testing period 2010-2019. Their associated training periods precede them and start in July 1963.

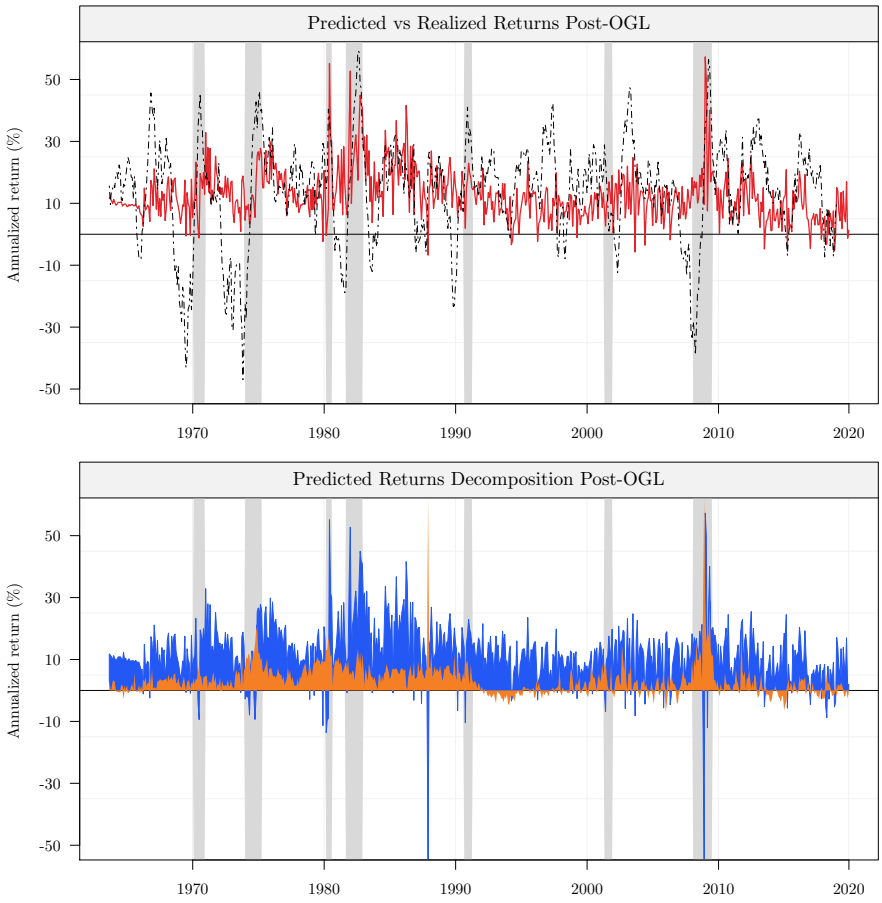

Figure 4: Predicted excess returns, realized excess returns, and prediction decomposition for the Fama-French five-factor model and an equally-weighted portfolio with the post-OGL method. In the upper panel, the predicted excess return path corresponds to the red plain line. The realized excess returns 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.

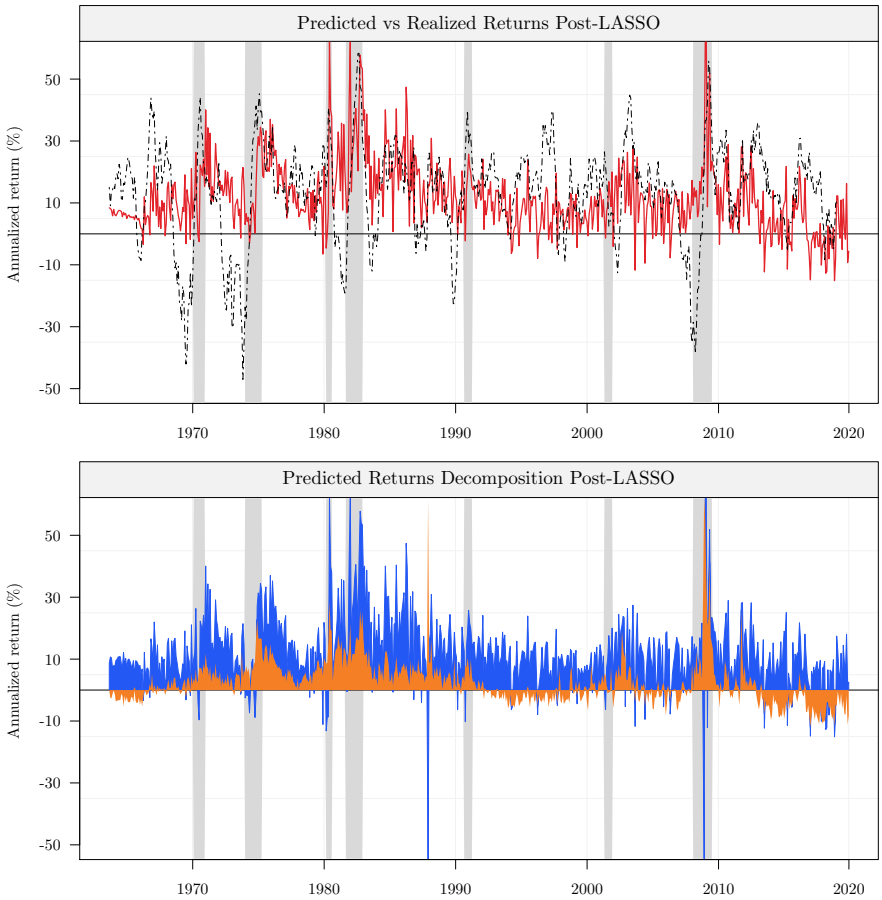

Figure 5: Predicted excess returns, realized excess returns, and prediction decomposition for the Fama-French five-factor model and an equally-weighted portfolio with the post-LASSO method. In the upper panel, the predicted excess return path corresponds to the red plain line. The realized excess returns 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.

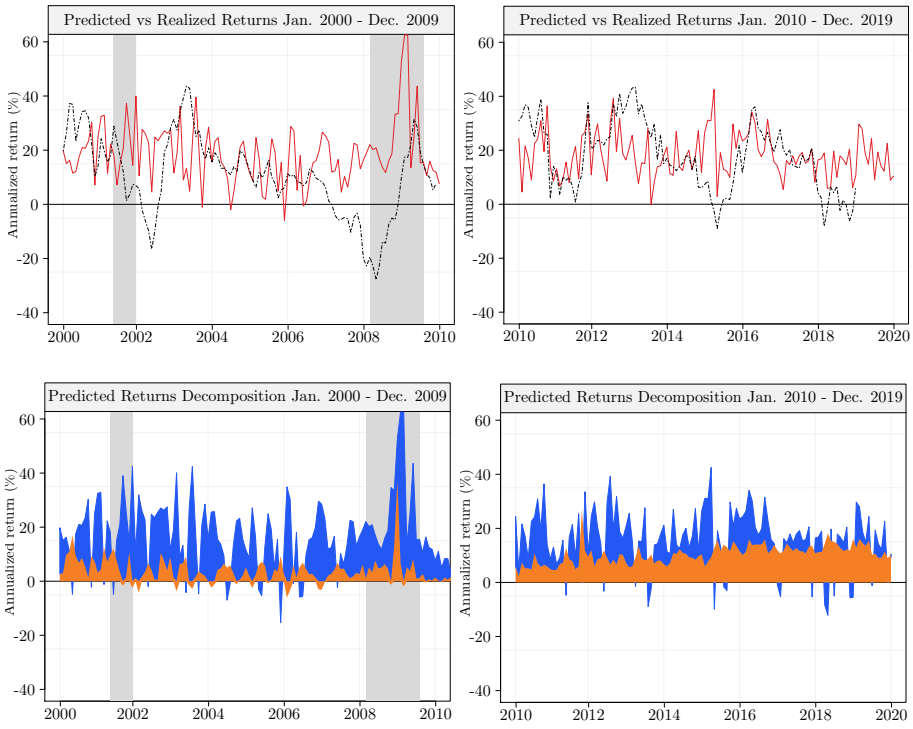

Figure 6: Out-of-sample predicted excess returns, realized excess returns, and prediction decomposition for the Fama-French five-factor model and an equally-weighted portfolio with the post-OGL method. The left panel is for the testing period 2000-2009. The right panel is for the testing period 2010-2019. Their associated training periods precede them and start in July 1963. 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).

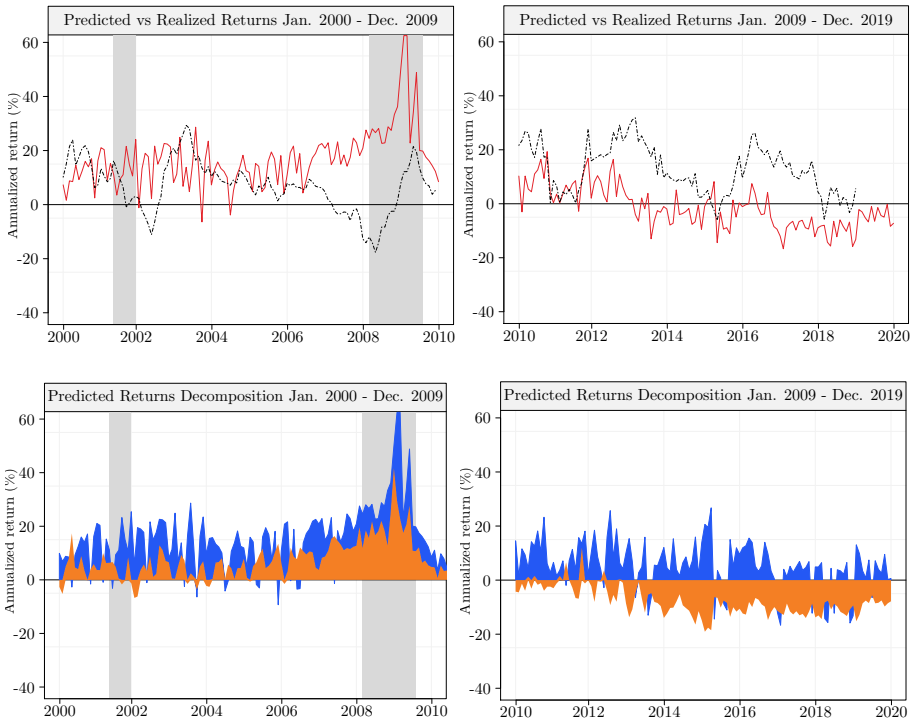

Figure 7: Out-of-sample predicted excess returns, realized excess returns, and prediction decomposition for the Fama-French five-factor model and an equally-weighted portfolio with the post-LASSO method. The left panel is for the testing period 2000-2009. The right panel is for the testing period 2010-2019. Their associated training periods precede them and start in July 1963. 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).

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

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. LSE manuscript.

Bryzgalova, S., Huang, J., and Julliard, C. (2019). Bayesian solutions for the factor zoo: We just ran two quadrillion models. Working Paper available on SSRN.

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

Chaieb, I., Langlois, H., and Scaillet, O. (2020). Factors and risk premia in individual international stock returns. Journal of Financial Economics, forthcoming.

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

Chernozhukov, V., Hansen, C., , and Spindler, M. (2015). Valid post-selection and post-regularization inference: An elementary, general approach. Annual Review of Economics, 7:649–688.

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

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.

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.

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.

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.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. Annals of Statistics, 28(5):1356–1378.

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

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.

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.

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, we need to define the following vector $x_t(\gamma) = (\text{vech}(X_t), Z_{t-1}^\top \otimes Z_{t-1}(\gamma)^\top, f_t^\top \otimes Z_{t-1}^\top, f_t^\top \otimes Z_{t-1}(\gamma)^\top)^\top$.

ASSUMPTION B.1: *There exist constants $\eta, \bar{\eta} \in (0,1]$ and $C_1, C_2, C_3, C_4 > 0$ such that for all $\delta > 0$ and $T \in \mathbb{N}$ we have:*

$$\sup_{\gamma \in [0,1]} \Pr\left[\left\| \frac{1}{T} \sum_t \left(x_t(\gamma) x_t(\gamma)^\top - \mathbb{E}\left[x_t(\gamma) x_t(\gamma)^\top \right] \right) \right\| \geq \delta \right] \leq C_1 T \times \exp\{-C_2 \delta^2 T^\eta\} + C_3 \delta^{-1} \exp\{-C_4 T^{\bar{\eta}}\}.$$

ASSUMPTION B.2: *There exists a constant M such that a)* $\sup_{\gamma \in [0,1]} \|x_t(\gamma)\| \leq M$, *P-a.s.. Moreover,*
b) $\sup_{\gamma \in [0,1]} \|A(\gamma)\| < \infty$, $\sup_{\gamma \in [0,1]} \|B(\gamma)\| < \infty$, $\sup_{\gamma \in [0,1]} \|C(\gamma)\| < \infty$.

ASSUMPTION B.3: $\inf_{\gamma \in [0,1]} \mathbb{E}[I_t(\gamma)] > 0$.

ASSUMPTION B.4: $\inf_{\gamma \in [0,1]} \text{eig}_{\min} \|\mathbb{E}[x_t(\gamma) x_t(\gamma)^\top]\| > 0$, *where* $\text{eig}_{\min}$ *denotes the minimum eigenvalue of* $\|\mathbb{E}[x_t(\gamma) x_t(\gamma)^\top]\|$.

ASSUMPTION B.5: *The trimming constants satisfy* $\chi_{1,T} = O\left((\log T)^{\kappa_1}\right)$, $\chi_{2,T} = O\left((\log T)^{\kappa_2}\right)$, *with* $\kappa_1, \kappa_2 > 0$.

ASSUMPTION B.6: *There exists a constant $M > 0$, such that* $\|\mathbb{E}[u_t, u_t^\top | Z_{t-1}]\| = \Sigma \otimes I_K \leq M$, *for all t, where* $u_t = f_t - \mathbb{E}[f_t | \mathcal{F}_{t-1}]$, *and* Σ, *the covariance matrix is diagonal.*

Assumption B.1 gives an upper bound for large-deviation probabilities of the sample average of random matrices $(x_t(\gamma) x_t(\gamma)^\top)$ uniformly w.r.t. $\gamma \in [0,1]$. It implies that the sample moments of squared components of the regressor vector converge in probability to the corresponding population moments at a rate $O(T^{-\eta/2} \log(T)^c)$, for some $c > 0$. Assumption B.2 eases the proofs and requires uniform upper bounds on the regressor values, intercept, and model coefficients. Assumption B.3 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.4 bounds away from zero the minimum eigenvalue of the population squared moment to exclude asymptotic multicolinearity problems uniformly across assets. Assumption B.5 gives an upper bound on the divergence rate of the trimming constants such that logarithmic divergence rate allows to control the post-OGL estimation error in the second-pass regression. Finally, Assumption B.6 bounds conditional variance-covariance matrix for the linear innovation u_t associated with the factor process and defines the matrix Σ as diagonal. This assumption helps to prove consistency of the LASSO estimator $\hat{F}_k$ equation per equation, when regressing $f_{k,t}$, for $k = 1, ..., K$, on Z_{t-1}.

B Proof of Proposition 1

In order to proof Proposition 1, we study the quantity $\hat{\nu}(\hat{\mathcal{S}})$. From the definition of $\hat{\nu}(\hat{S})$, we know that $\hat{\nu}(\hat{\mathcal{S}}) - \nu = \hat{Q}_{\beta_3}^{-1} \frac{1}{n} \hat{\beta}_{3,i}^{\top} \hat{w}_i C_{\nu,i}^{\top} (\hat{\beta}_i(\hat{\mathcal{S}}) - \check{\beta}_i)$. For $\hat{\nu}(\hat{\mathcal{S}})$ to be consistent, the support recovery $\hat{\mathcal{S}}$ needs to be consistent uniformly for β_i across all $i = 1, \ldots, n$, such that the following lemma holds.

LEMMA 1: *(Consistency $\hat{\beta}_i(\hat{\mathcal{S}})$)*
Under Assumptions A.4, A.6, B.1, Conditions C1 and C2 from Jacob et al. (2009) and Assumptions SC.1 and SC.2, we have $\sup_i \mathbf{1}_i^{\chi} \|\hat{\beta}_i(\hat{\mathcal{S}}) - \check{\beta}_i\| = \mathcal{O}_{p,\log}(T^{-\eta/2})$, *where the notation* $B_{n,T} = \mathcal{O}_{p,\log}(a_{n,T})$, *is such that* $B_{n,T}/a_{n,T}$ *is bounded in probability by some power of* $\log(T)$ *as* $n, T \to \infty$.

As mentioned above, to prove Lemma 1, we need uniform support recovery across all assets i. Therefore, we introduce the following lemma.

LEMMA 2: *(Uniform support recovery)*
Under Assumptions A.4, A.6, and Conditions C1 and C2 from Jacob et al. (2009), we have $\Pr(\mathcal{S} \subseteq \hat{\mathcal{S}}) \to 1$, *as* $n, T \to \infty$.

PROOF OF LEMMA 2: By Bonferroni inequality we have

$$\begin{aligned}\Pr\left(\mathcal{S} \subseteq \hat{\mathcal{S}}\right) &= \Pr\left(\mathcal{S}_1 \subseteq \hat{\mathcal{S}}_1, \ldots, \mathcal{S}_n \subseteq \hat{\mathcal{S}}_n\right) \geq 1 - \sum_{i=1}^{n} \Pr\left(\mathcal{S}_i \not\subseteq \hat{\mathcal{S}}_i\right) \\ &\geq 1 - n \max_{i=1,\ldots,n} \left\{\Pr\left(\mathcal{S}_i \not\subseteq \hat{\mathcal{S}}_i\right)\right\}.\end{aligned}$$

By Assumption A.6, we have $\Pr(\mathcal{S}_i \not\subseteq \hat{\mathcal{S}}_i) = \mathcal{O}(T^{-\omega})$, for all $i = 1, \ldots, n$. Hence, we obtain $\Pr(\mathcal{S} \subseteq \hat{\mathcal{S}}) = 1 - \mathcal{O}(T^{\bar{\gamma}-\omega}) = 1 - o(1)$, since $\bar{\gamma} < \omega$, by Assumption A.6, which concludes the proof. ■

Lemma 2 shows that the probability of recovering the true support $\mathcal{S}_i$ for all i tends to one P-a.s. Hence, from Lemma 2, the proof of Lemma 1 follows:

PROOF OF LEMMA 1: From Lemma 3 i) of GOS, under Assumption B.1, Assumptions SC.1 and SC.2 of GOS, we know that for $i = 1, \ldots, n$,

$$I_1 = \sup_i \mathbf{1}_i^{\chi} \|\hat{\beta}_i(\mathcal{S}_i) - \check{\beta}_i\| = \mathcal{O}_{p,\log}\left(T^{-1/2} \sup_i \|Y_{i,T}\|\right), \tag{20}$$

with $\hat{\beta}_i(\mathcal{S}_i)$ the estimator of $\check{\beta}_i$ under the true support $\mathcal{S}_i$ and $Y_{i,T} = 1/\sqrt{T} \sum_t I_{i,t} x_{i,t} \varepsilon_{i,t}$. In the framework of GOS, $\check{\beta}_i$ is given by β_i, as the model specification is not sparse for all i. Then, we can compute the following probability

$$\Pr(I_1) = \Pr\left(I_1 | \mathcal{S} \subseteq \hat{\mathcal{S}}\right)\left[1 - \Pr\left(\mathcal{S} \not\subseteq \hat{\mathcal{S}}\right)\right] + \Pr\left(I_1 | \mathcal{S} \not\subseteq \hat{\mathcal{S}}\right) \Pr\left(\mathcal{S} \not\subseteq \hat{\mathcal{S}}\right),$$

which, from the result in Lemma 2, can be written as

$$\begin{aligned}\Pr\left(I_1\right) &= \Pr\left(I_1|\mathcal{S}\subseteq\hat{\mathcal{S}}\right) + \left[-\Pr\left(I_1|\mathcal{S}\subseteq\hat{\mathcal{S}}\right) + \Pr\left(I_1|\mathcal{S}\not\subseteq\hat{\mathcal{S}}\right)\right]\Pr\left(\mathcal{S}\not\subseteq\hat{\mathcal{S}}\right)\\ &\leq \Pr\left(I_1|\mathcal{S}\subseteq\hat{\mathcal{S}}\right) + \Pr\left(\mathcal{S}\not\subseteq\hat{\mathcal{S}}\right) = \Pr\left(I_1|\mathcal{S}\subseteq\hat{\mathcal{S}}\right) + o(1).\end{aligned}$$

Hence, from (20), we can work conditionally on having selected the correct support for each asset i, so that lock $\sup_i \mathbf{1}_i^{\chi}\|\hat{\beta}_i(\hat{\mathcal{S}}) - \check{\beta}_i\| = \mathcal{O}_{p,\log}\left(T^{-1/2}\|Y_{i,T}\|\right)$. Moreover, from the result of Lemma 3 i) and its proof in GOS, and $\delta_T = T^{-\eta/2}(\log T)^{(1+\bar{\gamma})/(2C_2)}$, for $\eta, C_2, \bar{\gamma} > 0$, we obtain from Assumption B.1 that $\Pr(T^{-1/2}\sup_i\|Y_{i,T}\| \geq \delta_T) = \mathcal{O}(1)$, which concludes the proof.

■

Consistency of $\hat{\nu}(\hat{\mathcal{S}})$ follows from Lemma 1, and the following results ii) $\sup_i\|w_i\| = \mathcal{O}(1)$, iii) $1/n\sum_i\|\hat{w}_i - w_i\| = o_p(1)$ and iv) $\hat{Q}_{\beta_3} - Q_{\beta_3} = o_p(1)$ coming from Lemma 3 of GOS under Assumptions B.1, B.5 and B.6.