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

Olivier Scaillet

GFRI, University of Geneva
Swiss Finance Institute

joint work with
Gaetan Bakalli (Emlyon Business School)
Stéphane Guerrier (GSEM and Faculty of Science, University of Geneva)

October 27th, 2022
AI Big Data in Finance Research Forum Webinar

Context

- There is a recent burgeoning literature

  on using machine learning and **penalization** techniques

  in **factor** models,

  see e.g. the papers by the organizers of the ABFR forum

- Why? Many potential modeling choices and large cross-sections

  $\Longrightarrow$ need of a **structural approach to big data**
  based on finance theory for guidance.

Motivations

- Select models **compatible with finance theory** (No-arbitrage ex-ante).
- Understand which stock returns need **time-variation** in their factor loadings for prediction purpose.

	Carhart four-factor		Fama-French five-factor	
Methods	TI (%)	Arbitrage (%)	TI (%)	Arbitrage (%)
aOGL	38	0	35	0
aLASSO	46	100	31	100
time-invariant	100	0	100	0

Two-pass penalized methodology

$$R_{i,t} = a_{i,t} + b_{i,t}^{\top} f_t + \varepsilon_{i,t}.$$

No-arbitrage restriction: $a_{i,t} = b_{i,t}^{\top} \nu_t$

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

First-pass regression

- Adaptive Group-LASSO with Overlap (aOGL) to select and estimate $b_{i,t}$ drivers.

Second-pass regression

- Cross-sectional regression to estimate ν_t.
- Adaptive LASSO to select and estimate the driver components of $\mathbb{E}\left[f_t | \mathcal{F}_{t-1}\right]$.

Conditional linear factor model

From Gagliardini et al. (2016) under their Assumptions APR.1, APR.2 APR.3, SC.1 and SC.2, we consider the following time-varying factor model for assets $i = 1, \ldots, n$:

$$R_{i,t} = a_{i,t} + b_{i,t}^{\top} f_t + \varepsilon_{i,t}, \tag{1}$$

where $R_{i,t}$ denotes the excess return on asset i at period $t = 1, \ldots, T$, vector $f_t \in \mathbb{R}^K$ gathers the values of the observable factors at date t, $a_{i,t} \in \mathbb{R}$ and factor loadings $b_{i,t} \in \mathbb{R}^K$ are $\mathcal{F}_{t-1}$-measurable, where the filtration process $\mathcal{F}_{t-1}$ is the information available to all investors at time $t-1$.

- $\mathbb{E}[\varepsilon_{i,t} | \mathcal{F}_{t-1}] = 0$.
- $\text{Cov}(\varepsilon_{i,t}, f_{t,k} | \mathcal{F}_{t-1}) = 0, \ k = 1, ..., K$.

Asset pricing restriction

Under Assumptions APR.4 of Gagliardini et al. (2016), the following asset pricing restriction holds:

$$a_{i,t} = b_{i,t}^{\top} \nu_t, \tag{2}$$

where random vector $\nu_t \in \mathbb{R}^K$ is unique and is $\mathcal{F}_{t-1}$-measurable, which also takes the equivalent usual form:

$$\mathbb{E}\left[R_{i,t} | \mathcal{F}_{t-1}\right] = b_{i,t}^{\top} \lambda_t, \tag{3}$$

with $\lambda_t = \nu_t + \mathbb{E}[f_t | \mathcal{F}_{t-1}] \in \mathbb{R}^K$.

Conditioning information $\mathcal{F}_{t-1}$

The conditioning information $\mathcal{F}_{t-1}$ contains:

- $Z_{\underline{t-1}}$, where $Z_{t-1} \in \mathbb{R}^p$ is a vector of lagged instruments common to all stocks (i.e., inflation, stock variance, ...)
- $Z_{\underline{i,t-1}}$ where $Z_{i,t-1} \in \mathbb{R}^q$, is a vector of lagged characteristics specific to stock i (i.e., investment, past stock performance, ...)

where $Z_{\underline{t}} = \{Z_t, Z_{t-1}, ...\}$ denotes the set of past realizations.

Time-varying specification

Assumption A.1: (Sparse time-varying factor loadings)

$b_{i,t} = A_i + B_i Z_{t-1} + C_i Z_{i,t-1}$, where $A_i \in \mathbb{R}^K$ corresponds to a time-invariant model, and $B_i \in \mathbb{R}^{K \times p}$, $C_i \in \mathbb{R}^{K \times q}$ are sparse matrices of coefficients.

Assumption A.2: (Sparse time-varying risk premia)

(i) $\lambda_t = \Lambda_0 + \Lambda_1 Z_{t-1}$, where $\Lambda_0 \in \mathbb{R}^K$ corresponds to a time-invariant model and $\Lambda_1 \in \mathbb{R}^{K \times p}$ is a sparse matrix.
(ii) $\mathbb{E}\left[f_t | \mathcal{F}_{t-1}\right] = F_0 + F_1 Z_{t-1}$, where $F_0 \in \mathbb{R}^K$ corresponds to a time-invariant model and $F_1 \in \mathbb{R}^{K \times p}$ is a sparse matrix.

Assumption A.3: (Non sparse time-invariant contribution)

The time-invariant contribution is specified as $A_i^\top \left(\Lambda_0 - F_0\right) + A_i^\top f_t$. 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.

Data generating process

From (1), (2) and the above Assumptions, we get the following Data Generating Process (DGP),

$$\begin{aligned} R_{i,t} = & \underbrace{A_i^\top (\Lambda_0 - F_0)}_{\substack{\text{Time-invariant} \\ \text{intercept}}} + \underbrace{A_i^\top (\Lambda_1 - F_1) Z_{t-1} + Z_{t-1}^\top B_i^\top (\Lambda_0 - F_0)}_{\text{Linear form in } Z_{t-1}} \\ & + \underbrace{Z_{t-1}^\top B_i^\top (\Lambda_1 - F_1) Z_{t-1}}_{\text{Quadratic form in } Z_{t-1}} \\ & + \underbrace{Z_{i,t-1}^\top C_i^\top (\Lambda_0 - F_0) + Z_{i,t-1}^\top C_i^\top (\Lambda_1 - F_1) Z_{t-1}}_{\text{Quadratic form in } Z_{t-1} \& Z_{i,t-1}} \\ & + \underbrace{A_i^\top f_t}_{\substack{\text{Time-invariant} \\ \text{factor loadings}}} + \underbrace{Z_{t-1}^\top B_i^\top f_t + Z_{i,t-1}^\top C_i^\top f_t}_{\text{Bi-linear form in } f_t} + \varepsilon_{i,t}. \end{aligned} \tag{4}$$

No-arbitrage restrictions at stock level

To create a model compatible with the no-arbitrage restriction, we have the following constraints:

Model construction constraints

1. If the selection procedure selects **at least one variables** in Z_{t-1} $(Z_{i,t-1})$ in the intercept $a_{i,t}$, then it should also select **at least one element** in Z_{t-1} $(Z_{i,t-1})$ in $b_{i,t}$.
2. If the selection procedure selects **no variables** in Z_{t-1} $(Z_{i,t-1})$ in the intercept $a_{i,t}$, than it should also select **no elements** in Z_{t-1} $(Z_{i,t-1})$ in $b_{i,t}$.

To implement estimation and selection procedure on (4)
$\Longrightarrow$ we redefine the regressors to obtain a generic panel model.

Generic panel model

Definition

- $\tilde{Z}_{t-1} = (1, Z_{t-1}^\top)^\top \in \mathbb{R}^{\tilde{p}}$, where $\tilde{p} = p + 1$
- $\breve{B}_i = [A_i | B_i] \in \mathbb{R}^{K \times \tilde{p}}$.
- $\Lambda - F = [(\Lambda_0 - F_0)|(\Lambda_1 - F_1)] \in \mathbb{R}^{K \times \tilde{p}}$.

Hence we get the linear transformed regressors

$$x_{1,i,t} = \left(\text{vech}\,(X_t)^\top, \tilde{Z}_{t-1}^\top \otimes Z_{i,t-1}^\top\right)^\top \in \mathbb{R}^{d_1}, \quad d_2 = K\tilde{p} + Kq$$

$$x_{2,i,t} = \left(f_t^\top \otimes \tilde{Z}_{t-1}^\top, f_t^\top \otimes Z_{i,t-1}^\top\right)^\top \in \mathbb{R}^{d_2}, \quad d_1 = (\tilde{p}+1)\tilde{p}/2 + \tilde{p}q$$

with $x_{i,t} = (x_{1,i,t}^\top, x_{2,i,t}^\top)^\top$ of dimension $d = d_1 + d_2$, and $\beta_i = (\beta_{1,i}^\top, \beta_{2,i}^\top)^\top$, we get the following linear factor model in transformed parameter

$$R_{i,t} = \beta_i^\top x_{i,t} + \varepsilon_{i,t}. \quad (5)$$

Penalized two-pass regression

To implement a two-pass penalized regression on (1), (2) and (3), we:

Penalized first-pass regression

- Select and estimates the β_i through the aOGL procedure on (5) (Percival, 2012).

Penalized second-pass regression

- Compute the WLS estimator for the ν_t (Gagliardini et al., 2016).
- Select and estimate the matrix F of coefficients through the aLASSO estimator (Zou, 2006).

Why aOGL penalization?

Example

Let us study a simple example with $K = 2$ factors, $p = 1$ common instrument, and $q = 1$ characteristic, so that the regressors $x_{i,t} = (x_{1,i,t}^\top, x_{2,i,t}^\top)^\top$ become

$$x_{1,i,t} = (\underbrace{1}_{x_{1,i,t,1}}, \underbrace{2Z_{t-1}}_{x_{1,i,t,2}}, \underbrace{Z_{t-1}^2}_{x_{1,i,t,3}}, \underbrace{Z_{i,t-1}}_{x_{1,i,t,4}}, \underbrace{Z_{t-1}Z_{i,t-1}}_{x_{1,i,t,5}})^\top \in \mathbb{R}^5$$

and

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

Groups that satisfy no arbitrage *ex-ante*

Building on the above example, we need to create the following 6 groups to satisfy the no-arbitrage restriction:

- Time-invariant contribution group: $(x_{1,i,t,1}, x_{2,i,t,1}, x_{2,i,t,3})^\top$.
- Cross-product in $\tilde{Z}_{t-1}$: $(x_{1,i,t,2})$.
- Groups in scaled factors: $(x_{1,i,t,3}, x_{2,i,t,2})^\top$ and $(x_{1,i,t,3}, x_{2,i,t,4})^\top$.
- Groups with characteristics: $(x_{1,i,t,4}, x_{1,i,t,5}, x_{2,i,t,5})^\top$ and $(x_{1,i,t,4}, x_{1,i,t,5}, x_{2,i,t,6})^\top$.

Set of 32 possible models without arbitrage *ex-ante*

	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	$x_{1,5}$	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	$x_{2,5}$	$x_{2,6}$
$\mathcal{M}_1$	✓	✗	✗	✗	✗	✓	✗	✓	✗	✗	✗
$\mathcal{M}_2$	✓	✓	✗	✗	✗	✓	✗	✓	✗	✗	✗
$\mathcal{M}_3$	✓	✗	✓	✗	✗	✓	✓	✓	✗	✗	✗
$\mathcal{M}_4$	✓	✗	✓	✗	✗	✓	✗	✓	✓	✗	✗
$\mathcal{M}_5$	✓	✓	✓	✗	✗	✓	✓	✓	✗	✗	✗
$\mathcal{M}_6$	✓	✓	✓	✗	✗	✓	✗	✓	✓	✗	✗
$\mathcal{M}_7$	✓	✗	✓	✗	✗	✓	✓	✓	✓	✗	✗
$\mathcal{M}_8$	✓	✓	✓	✗	✗	✓	✓	✓	✓	✗	✗
$\mathcal{M}_9$	✓	✗	✗	✓	✓	✓	✗	✓	✗	✓	✗
$\mathcal{M}_{10}$	✓	✗	✗	✓	✓	✓	✗	✓	✗	✗	✓
$\mathcal{M}_{11}$	✓	✗	✗	✓	✓	✓	✗	✓	✗	✓	✓
$\mathcal{M}_{12}$	✓	✓	✗	✓	✓	✓	✗	✓	✗	✓	✗

Why aOGL penalization?

LASSO-type method

LASSO penalization **might select a model introducing arbitrage *ex-ante*** by

- Removing the scaled factors $x_{2,i,t,2} = Z_{t-1} f_{t,1}$ and $x_{2,i,t,4} = Z_{t-1} f_{t,2}$ from the full model.
- Removing the quadratic term $x_{1,i,t,3} = Z_{t-1}^2$, while keeping its corresponding scaled factors $x_{2,i,t,2}$ and $x_{2,i,t,4}$ from the full model.

Group-LASSO method

Group-LASSO penalization (Yuan and Lin, 2006) removes all groups where one covariate is included if one of those groups is shrunk towards zero (Jacob et al., 2009). Hence the procedure selects **either all scaled factors or none of them**.

Solution: The aOGL method with variables duplication.

Benefits of relying on Finance Theory

By relying on finance theory in our selection approach (no introduction of arbitrage *ex-ante*), we **reduce the space of possible models** under the LASSO approach through adequate grouping (aOGL), and obtain the following upper bound

$$\frac{2^{J-1}}{2^{d-n_1}} \leq \frac{1}{8},$$

where J is the number of groups and n_1, resp. n, the number of variables in the time-invariant group (K factors + intercept), resp. full model.

Usually, $\boldsymbol{J \ll d}$.
In our empirical application, we have $K = 5$ factors, $p = 6$ instruments and $q = 13$ characteristics ($d = 219$), and the ratio $\frac{2^{J-1}}{2^{d-n_1}}$ becomes $\mathbf{2^{-97}}$!!

First-pass estimation of $\hat{\beta}_i$

Considering the aOGL estimator for the factor loadings, we get the following result:

Lemma

We have that

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

where the vector V_i has entries

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

where $Q_{H_i,x,i}$ is the submatrix of $Q_{x,i} = \mathbb{E}[x_{i,t}x_{i,t}^\top|\gamma_i]$ with indices in H_i, and $H_i = \{l \in \{1,\ldots,d\} : \beta_{i,l} \neq 0\}$.

Second-pass estimation of ν

The parameter restriction in (2) can be re-written as

$$\beta_{1,i} = \beta_{3,i}\nu, \quad \nu = \text{vec}(\Lambda^\top - F^\top),$$

where we estimate ν with the following WLS estimator:

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

where $\hat{Q}_{\beta_3} = \frac{1}{n} \sum_i \hat{\beta}_{3,i}^\top \hat{w}_i \hat{\beta}_{3,i})$, and weights are estimates of $w_i = \mathbf{1}_i^\chi \left(\text{diag}\left[v_i\right]\right)^{-1}$, with v_i the asymptotic variances of the standardized errors.

Proposition 1: Consistency of $\hat{\nu}$

Under Assumptions APR.1 to APR.4, SC.1 and SC.2 of Gagliardini et al. (2016) and Assumptions A.1 to A.2, A.4 to A.7, we have that $\|\hat{\nu} - \nu\| = o_p\left(1\right)$, when $n, T \to \infty$.

Selection and estimation of F

We consider the aLASSO estimator of Zou (2006) to select and estimate the matrix F of coefficients. We get the final estimates of the sparse matrix Λ from the relationship $\text{vec}(\hat{\Lambda}^\top) = \hat{\nu} + \text{vec}(\hat{F}^\top)$, which yields $\hat{\lambda}_t = \hat{\Lambda} Z_{t-1}$.

Proposition 2: Consistency of $\hat{\Lambda}$

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

Data description

Base assets: 6874 stocks with monthly returns from Jul. 1963 to Dec. 2019 after merging CRSP and Compustat databases.

Sets of factors:

- $K = 4$ factors of Carhart (1997), $f_t = (f_{m,t}, f_{hml,t}, f_{smb,t}, f_{mom,t})^\top$.
- $K = 5$ factors of Fama and French (2015), $f_t = (f_{m,t}, f_{hml,t}, f_{smb,t}, f_{rmw,t}, f_{cma,t})^\top$.

Common instruments $p = 6$:

- Dividend yield
- Net equity expansion
- Inflation
- Stock variance
- Default spread
- Term spread

Data description

Characteristics $q = 13$ (see Freyberger et al. (2020) for details) :

- Change in share outstanding
- Log change in the split adjusted shares outstanding
- Growth rate in total assets
- Size
- Last month volume over shares outstanding
- Adjusted profit margin
- Momentum
- Intermediate momentum
- Short-term reversal
- Closeness to 52-week high
- Ratio of market value of equity plus long-term debt minus total assets to Cash and Short-Term Investments
- Standard unexplained volume
- Total volume

First-pass regression results

	Carhart four-factor		Fama-French five-factor	
Methods	TI (%)	Arbitrage (%)	TI (%)	Arbitrage (%)
aOGL	38	0	35	0
aLASSO	46	100	31	100
time-invariant	100	0	100	0

Messages

- **100%** of models selected trough aLASSO estimator introduce **arbitrage *ex-ante***.
- **Approximately 2/3** of the assets need time-variation in their factor loadings for prediction purpose.

Selection results Fama-French five-factor

Messages

- In line with Chaieb et al. (2020), the characteristics are **not necessarily paired more often with their corresponding factors.**
- It seems that the **common instruments Z_{t-1} are key drivers** of the time variation of the factor loadings.

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

Selection results Fama-French five-factor

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

Messages

The longer the sample size, the more "action" is needed for the dynamics of the factor loadings.

In-sample results: Prediction Error

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

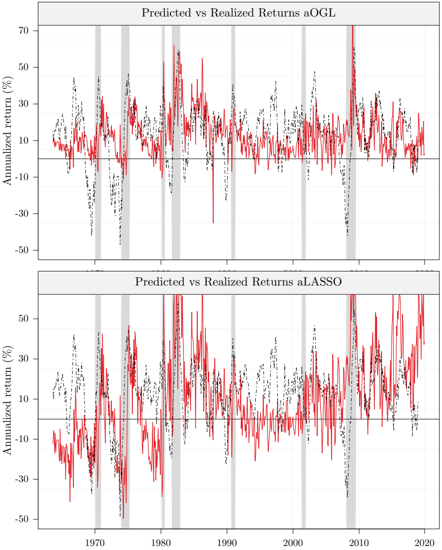
Messages

- PE computed on **an equally-weighted portfolio** for both factor specifications.
- **aOGL performs better than the aLASSO,** and comparable to $\lambda_t \& \nu$.

In-sample results: realized vs predicted returns (FF5)

Messages

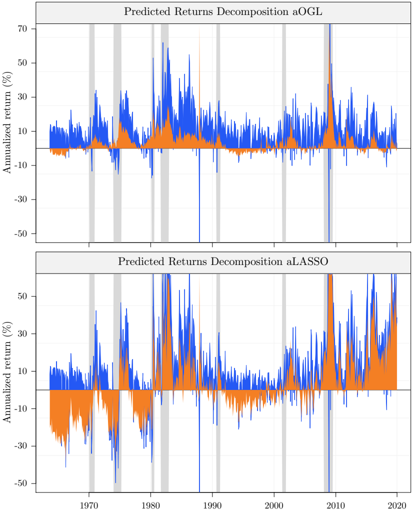
- aOGL predicted excess return paths (red plain line) **reconcile well** with the realized excess returns (black dashed line).
- aLASSO method sometimes predicts **large negative excess returns**: at odd with a positive reward expected from taking risks.



In-sample results: predicted returns decomposition (FF5): $\mathbb{E}\left[R_{i,t}|\mathcal{F}_{t-1}\right] = a_{i,t} + b_{i,t}^{\top}\mathbb{E}\left[f_t|\mathcal{F}_{t-1}\right]$

Messages

- aLASSO penalization ends up with time-varying models presenting arbitrage $\Longrightarrow$ **larger values for estimated** $\hat{a}_{i,t}$ (often in NBER recessions).
- The estimated path for $a_{i,t}$ is close to zero with the aOGL method (tradable factors).



Out-of-sample results: Prediction Error

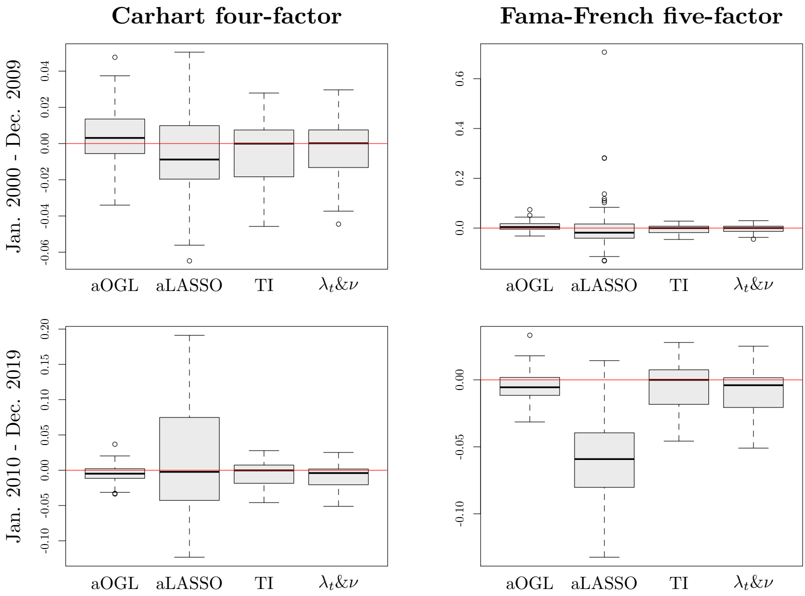
	Carhart four-factor					
	Jan. 2000 to Dec. 2009			Jan. 2010 to Dec. 2019		
Methods	RMSPE	Av(|PE|)	Std(|PE|)	RMSPE	Av(|PE|)	Std(|PE|)
aOGL	$1.58 \cdot 10^{-2}$	$1.23 \cdot 10^{-2}$	$1.00 \cdot 10^{-2}$	$1.34 \cdot 10^{-2}$	$1.06 \cdot 10^{-2}$	$0.83 \cdot 10^{-2}$
aLASSO	$2.43 \cdot 10^{-2}$	$2.03 \cdot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

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

Message

aOGL performs better than the aLASSO, and comparable to $\lambda_t\&\nu$.

Out-of-sample results: Distribution of PE



Out-of-sample R^2 Fama-French five-factor model

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

Message

aOGL performs better than the aLASSO, but advantage less clear compared to a time-invariant method.

Conclusions

Take-home messages

- Our empirical results show that taking explicitly into account the **no-arbitrage restriction** coming from the Arbitrage Pricing Theory does **help** in predictive modeling of large cross-sectional equity data sets with penalization methods.
- A **structural approach** to big data where we incorporate finance theory **improves** on the prediction performance of the estimated quantities.

Further applications

- A better predictive performance of excess returns should help to better gauge time-variation in the risk-reward trade-off.
- It should help to improve performance of time-varying portfolio allocation when we use predicted excess returns as inputs.

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.

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

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.

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.

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

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

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

Simulation study 1

We compare the selection and prediction performance for the aLASSO and aOGL under the following sparse setting.

- $B = 500$ replicates from the DGP in (4) of length $T_i = 500$.
- Error terms $\varepsilon_{i,t} \sim \mathcal{N}(0, \sigma^2)$, where $\sigma = 0.09$.
- $K = 5$ factors, $p = 6$ instruments, $q = 13$ characteristics, which yields $d = 219$ covariates.
- Number of non-zero coefficients is 28 (very sparse design).

Simulation study 1

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

Message

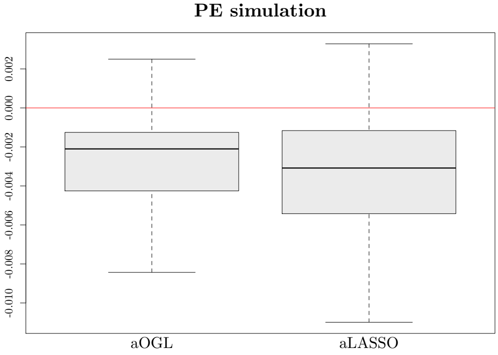
- **Average RMSPE lower for the aOGL method** compared to the aLASSO method.
- aLASSO based methods select **almost always models with arbitrage *ex-ante***.

Simulation study 2

We compare the prediction performance for the following 2 methods: post-LASSO, post-OGL at the porfolio level.

- Select randomly $n = 500$ assets from empirical data.
- Error terms ε_t is Gaussian with mean 0 and block (of length 50) diagonal correlation matrix such that $\text{corr}(\varepsilon_{k,t}, \varepsilon_{l,t}) = 0.25^{|k-l|}$, $k, l = 1, ..., 50, l \neq k$ and $\text{var}(\varepsilon_{i,t}) = 0.05$.
- Simulate from the DGP in (4) under the selected support from the OGL.
- Apply the post-OGL and post-LASSO estimators, with the WLS estimator of ν for each method.
- Investigate out-of-sample Prediction Errors (PE) for an equally-weighted portfolio.

Simulation study 2 : Prediction Error distributions



Message

- Post-OGL method PE are **closer to zero**.
- Post-OGL method exhibits a **narrower boxplot**.

Simulation study 2

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

Message

- **Lower average (RMSPE)** for aOGL method.
- **Lower average (MAPE)** for aOGL method.