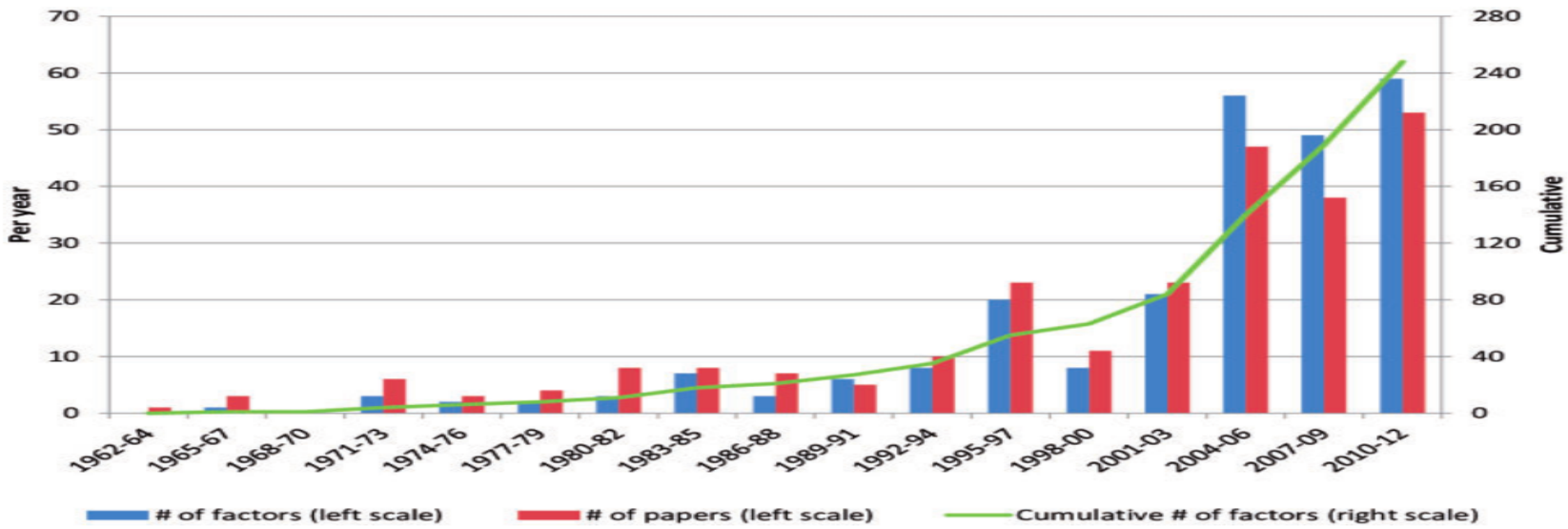
Regularising the factor zoo using OWL

Chuanping Sun

QMUL

Motivation

- Harvey, Liu and Zhu (2015)



- Cochrane (2011) in his presidential address: In the zoo of new variables, I suspect we will have to use different methods (portfolio sorting).

Literature

- McLean and Pontiff (2016): anomalies declined sharply after publication.
- Harvey et al.(2015): documented 316 factors; many as result of data-snooping.
- Green et al.(2016): use multiple test accounting for data-snooping to find factors for the US stock market.
- Fama, French (2016): RHS method and sharpe ratio to "choose factors" according to Barillas, Shanken (2016).
- Harvey and Liu (2017): "Lucky factors" $\rightarrow$ use orthogonal design and Bootstrap to find significant factors.
- Pelger and Lettau(2017):Risk premium PCA to estimate asset pricing factors
- Pukthuanthong et al.(2017): proposed a protocol to screen factors $\rightarrow$ factors must be correlated with test asset returns.
- Feng et al.(2017): "Taming the factor zoo"; two step lasso plus OLS post-selection regression to find cross-sectional return predictors.
- Ando and Bai (2014): use SCAD (smoothly clipped absolute deviation) to find Chinese stock predictors.
- Nagel et al. (2017): use elastic net (ℓ_1 plus ℓ_2 norm) to shrink the cross section in a Bayesian framework.
- Freyberger et al. (2017): non-parametric adaptive group lasso to find which characteristics provide independent information for the cross-sectional returns.
- Bryzgalova (2016): modified adaptive lasso in the Fama-MacBeth regression to shrink spurious factors.

using machine learning techniques to reduce high-dimensionality problems in finance.

Contributions

1. Utilise the Ordered and Weighted ℓ_1 norm regulariser (OWL) in machine learning literature to reduce high dimensionality in the "zoo of factors".
2. ! OWL relaxes orthogonal matrix design assumption (allow factors to be highly correlated). example
3. It answers two questions:
   - Which factors are redundant and weak factors in terms of explaining the cross section of average returns?
   - Which factors share the similarity in term of explaining the cross-sectional expected returns? (factors that are correlated and have similar explanatory power)

Contributions contd'

4. Two-Stage select-and-test procedure to find factors.
   - First stage, we use the OWL to shrink the high dimensionality of factors. Survival factors are grouped by their magnitude (clustered factors).
   - Second stage, group-wise orthogonal test for factor significance. $\Rightarrow$ which factors provide independent information about average returns? Cochrane (2011).

SDF

- $m_t = 1 - b(f_t - \mu_f)$: demeaned and normalised SDF
- The Euler equation states: $E[R_t^e m_t] = 0$, $\forall$ admissible SDF $m_t \in \mathbb{M}$. For a candidate $m_t(b)$ where b are the model parameters yet to be estimated, the pricing error $e(b) = E[R_t^e m_t(b)]$.
- With the specification of m_t, we can write:

$$\begin{aligned} e(b) &= E[R_t m_t(b)] = E[R_t]E[m(b)] + cov(R_t, m(b)) \\ &= \mu_{R^e} - Cb \end{aligned}$$

Where C is the covariance matrix of returns and factors, $C = cov(R^e, f)$

Recover the model by minimising the discrepancy

$$\hat{b} = \underset{b}{argmin} \;\; e(b)' * W * e(b)$$

- By choosing different W, we can arrive different measures of discrepancy. The most well known choice of W is the GMM optimal weighting matrix, that is, the inverse of variance matrix, however it would be incorrect in the context of comparing models.
- A popular choice of W can be the identity matrix, which avoids favouring the more volatile assets (Ludvigson, 2012).
- If choose $W = E(RR')^{-1}$, the discrepancy measure would corresponds to the well known Hansen-Jaganathan distance.

Recover the model by minimising the discrepancy

Using the standard GMM method, we can estimate

$$\hat{b} = \underset{b}{argmin}\ \ e(b)' * W * e(b) = (C'WC)^{-1}C'W\mu_{R^e}$$

- Ludvigson(2012) advocates to use the Identity matrix as the weighting matrix when the test assets are decided, as it would yield more stable result comparing using an estimated weighting matrix.

!!! The curse of dimensionality: When the dimension of C is big, ($N \not\gg K$, or even $N < K$), the traditional method will fall in short. It cries out for regularisation.

Ordered and Weighted ℓ_1 norm (OWL) regulariser

Proposition 1

$$\hat{b} = \underset{b}{argmin} \frac{1}{2}(\mu_R - Cb)' W_T (\mu_R - Cb) + \Omega_\omega(b) \tag{1}$$

[a] *where* $\Omega_\omega(b) = \omega' |b|_\downarrow$ *, and* ω *is a* $K \times 1$ *vector, and* $\omega \in \kappa$*, where* κ *is a monotone non-negative cone, defined as* $\kappa := \{x \in R^n : x_1 \geq x_2 \geq ... \geq x_n \geq 0\}$ *and* $\omega_1 > \omega_K$*.* $|b|_\downarrow$ *is the absolute value of the parameter, decreasingly ordered by its magnitude.*

[a] For the ease of notation, I will use μ_R to denote the mean of EXCESS returns, without explicitly using the e subscript.

Atomic Norm of $\Omega_\omega(.)$

Figure: Atomic Norm in R^3, Figueiredo et al.(2015)

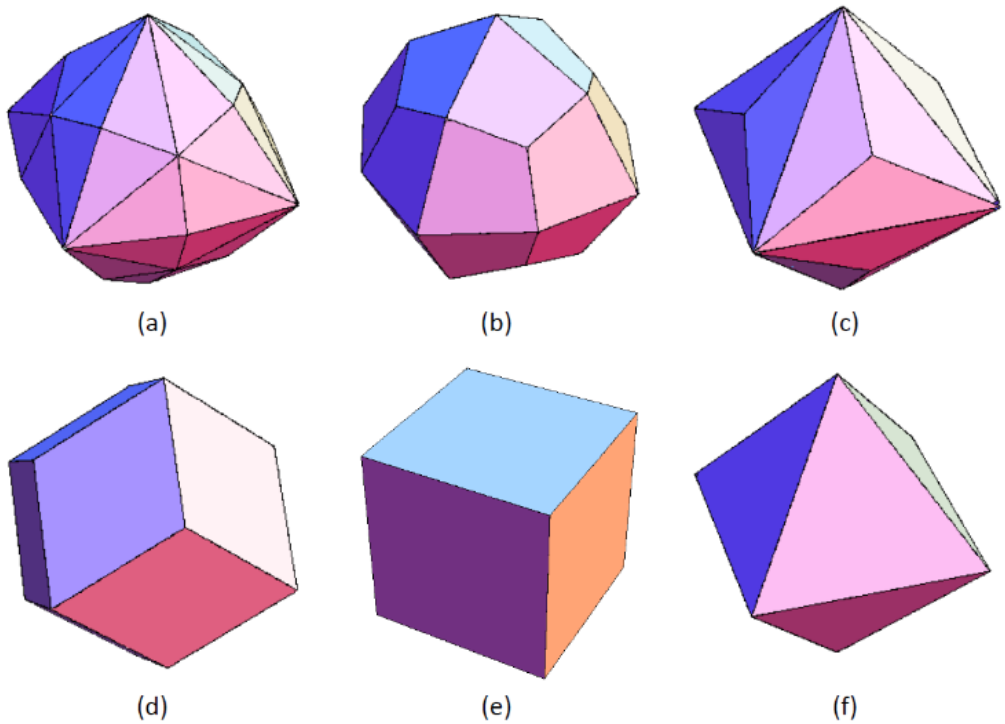


Fig. 2. OWL balls in $\mathbb{R}^3$ with different weights: (a) $w_1 > w_2 > w_3 > 0$; (b) $w_1 > w_2 = w_3 > 0$; (c) $w_1 = w_2 > w_3 > 0$; (d) $w_1 = w_2 > w_3 = 0$; (e) $w_1 > w_2 = w_3 = 0$; (f) $w_1 = w_2 = w_3 > 0$.

Proximal algorithm

define the proximal function as

$$Prox_{\Omega_\omega}(b) = argmin_x \frac{1}{2}||x-b||_2^2 + \Omega_\omega(x) \tag{2}$$

since $\Omega_\omega(b) = \Omega_\omega(|b|)$, and $||b - sign(b) \odot |x|||_2^2 \leq ||b-x||_2^2$, we have:

$$Prox_{\Omega_\omega}(b) = sign(b) \odot Prox_{\Omega_\omega}(|b|) \tag{3}$$

Now since $\Omega_\omega(x) = \Omega_\omega(Px)$ and $||P(v-x)||_2^2 = ||v-x||_2^2$ where P is a permutation matrix. we have:

$$Prox_{\Omega_\omega}(b) = sign(b) \odot Prox_{\Omega_\omega} P'(|b|)\Omega_\omega(|b|_\downarrow) \tag{4}$$

where $|b|_\downarrow$ is decreasingly ordered absolute value of coefficients. and $P'(|b|)$ is the transpose of the permutation matrix, which recovers the order.

Proximal algorithm contd'

For any $b \in \kappa$, we have:

$$\begin{aligned}\frac{1}{2}||x-b||_2^2 + \Omega_\omega(x) &= \frac{1}{2}||x||_2^2 + \frac{1}{2}||b||_2^2 - b'x + \Omega_\omega(x) \\ &\geq \frac{1}{2}||x^*||_2^2 + \frac{1}{2}||b||_2^2 - b'x^* + \Omega_\omega(x^*)\end{aligned}$$

where $x^* \in \kappa$. So $Prox_{\Omega_\omega}(b) \in \kappa$, and $\Omega_\omega(x) = \omega' x$, then we have:

$$argmin_{x\in\kappa}\frac{1}{2}||x-|b|_\downarrow||_2^2 + \omega' x = argmin_{x\in\kappa}\frac{1}{2}||x-(|b|_\downarrow - \omega)||_2^2$$

which is the projection of $(|b|_\downarrow - \omega)$ onto κ, Then equation (4) can be written as:

$$Prox_{\Omega_\omega}(b) = sign(b) \odot (P'(|b|)Proj_\kappa(|b|_\downarrow - \omega)) \tag{5}$$

where $Proj_\kappa(.)$ is the projection operator onto κ.

FISTA (fast interactive soft-thresholding)for OWL

Input: μ_R, C, ω

Output: $\hat{b}$ **in (1)**

Initialisation: $b_0 = \hat{b}_{OLS}, t_0 = t_1 = 1, u_1 = b_0, k = 1, \eta \in (0,1), \tau_0 \in (0, 1/L)$

while *some stopping criterion not met* **do**

  $\tau_k = \tau_{k-1}$;

  $b_k = Prox_{\Omega_\omega}(u_k + \tau * C' * (\mu_R - Cb))$

  while $||\mu_R - Cb||_2^2 > Q(b_k, u_k)$ **do**

    $\tau_k = \eta * \tau_k$;

    $b_k = Prox_{\Omega_\omega}(u_k + \tau * C' * (\mu_R - Cb))$

  end

  $t_{k+1} = (1 + \sqrt{1 + 4t_k^2})/2$

  $u_{k+1} = b_k + \frac{t_{k-1}}{t_{k+1}}(b_k - b_{k-1})$

  $k \leftarrow k + 1$

end

Return: b_{k-1}

Tuning parameters

- Although by choosing different weighting scheme, we can arrive different LASSO norm specification, we restrict our weighting scheme consistent with OSCAR (octagonal selection and clustering algorithm for regression) because of its clustering property, that is linear and equal-spaced. In OSCAR the weighting vector can be specified by two tuning parameters, λ_1 and λ_2: $\omega_i = \lambda_1 + (K - i)\lambda_2$, where K is the total number of factors in the model, and $i = 1, 2, ..., K$.

Cross-validation

- We use the common 5-fold cross validation method, that is, given grids of λ_1 and λ_2, at each point on the grids (mesh), we estimate the model using OWL. In particular, we divide the sample into 5 parts, using 4 parts to estimate the model using OWL, and use 1 part to estimate the out-of-sample estimation error (MSE), we rotate these parts as being used as the out-of-sample sub-sample, and then compute the average MSE. At each point on the mesh, we compute the MSE, and then we compare all MSEs obtained at different points on the mesh. The one with the smallest MSE would corresponds to the optimal parameters.

Statistical Properties

Theorem 1 (Error bounds)

Let the DGP be $\mu_R = Cb^* + e$, $b^* \in R^K$ *is* $S-$*sparse,* $e \in R^N$ *is the error term, and* $||e||_1/n \leq \epsilon$. *Let* $\hat{b}$ *be a solution of (1),* ω_1 *is the first element of the weighting vector* ω, *and* $\bar{\omega}$ *is the mean of all elements of* ω, *then*

$$E||\hat{b} - b^*||_2 = \mathcal{O}(||b^*||_2 \frac{\omega_1}{\bar{\omega}} \sqrt{\frac{SlogK}{N}})$$

Statistical Properties

Comments:

- OWL is a biased estimator.
- OWL convergence rate is of $\sqrt{\frac{SlogK}{N}}$.
- estimation bias is proportional to weights.
- OWL shrinks more of parameters when its true (absolute) value is great, shrinks less of parameters of small magnitude.

Statistical Properties

Theorem 2 (Grouping)

Let $\hat{b}(K \times 1)$ be a solution of (1), f_i and f_j (both $T \times 1$)be the ith and jth factors, so b_i and b_j are the coefficients in the SDF specification associated with the i^{th} and j^{th} factors. Let $\mu_R(N \times 1)$ be a vector of test asset means, and Δ_ω be the smallest distance between two successive weights in ω, if

$$\sigma_{f_i - f_j} < \frac{\Delta_\omega}{||\mu_R||_2 ||\sigma_R||_2}$$

then $\hat{b}_i = \hat{b}_j$.

Statistical Properties

Comments:

- When two factors are highly correlated $\rightarrow$ they are grouped together (having the same coefficients)
- The greater Δ (λ_2 in the OSCAR setting) $\rightarrow$ more grouping $\rightarrow$ because the atomic norm has more pointed surface $\rightarrow$ tangent point with the contour from the unregularised solutions.
- Less volatile of tests assets, more grouping $\rightarrow$ When portfolios returns are not much different from each other, factors are having less explanatory power.
- Smaller test asset returns, more grouping $\rightarrow$ When returns are very close to zero, most factors would be grouped together because of less explanatory power in all factors.

Two-stage group-wise testing procedure

1 The first stage, we obtain a sparse model from OWL, then group these factors according to their coefficient magnitude estimated through OWL in descending order; that is forming a sequential of groups $\{gp_1, gp_2, ..., gp_s\}$, in each group, it contains one or more factors. The elements in each group is equal to each other in terms of absolute value. So the coefficient in gp_i is greater than in gp_j, $\forall j > i$.

2.1 The second stage is orthogonal regression. First, regress μ_R on first group (gp_1) of C, find all significant factors, and include them in active set $\mathbb{A}$.

Two-stage group-wise testing procedure

2.2 Regress μ_R on the updated active set $\mathbb{A}$, obtain residual vector $\mathbb{V}$, and regress $\mathbb{V}$ on the next group of C, test for significance, and then update $\mathbb{A}$ by including more newly tested (significant) factors.

2.3 Repeat step 2.2, until no more significant factors are found in a new group, or all groups have been explored. Then the tested model would be the factors included in the active set $\mathbb{A}$.

Why need the 2^{nd} stage?

- OWL is a biased estimator, the estimation error is proportional to the parameter's true value. Which means less shrinkage to weak factors $\rightarrow$ possible spurious factors.
- The grouping property nicely classifies factors by their magnitudes $\rightarrow$ we can test groups step by step $\rightarrow$ less prone to spurious factor issues.
- we use orthogonal transformation after testing each group, that makes each group of factors provides independent information to the cross-sectional of returns.

Simulation design

Simulate a return-factor covariance matrix C ($N \times K$), with following correlation structure.
Let ρ ($K \times K$ matrix) denotes the correlation matrix of C,
$\rho_1, \rho_2, \rho_3 \in (-1, 1)$ and ρ are divided into 3 blocks such that:

$$bk_1 = \begin{pmatrix} 1 & \dots & \rho_1 \\ \vdots & \ddots & \vdots \\ \rho_1 & \dots & 1 \end{pmatrix}; bk_2 = \begin{pmatrix} 1 & \dots & \rho_2 \\ \vdots & \ddots & \vdots \\ \rho_2 & \dots & 1 \end{pmatrix}; bk_3 = \begin{pmatrix} 1 & \dots & \rho_3 \\ \vdots & \ddots & \vdots \\ \rho_3 & \dots & 1 \end{pmatrix}$$

and

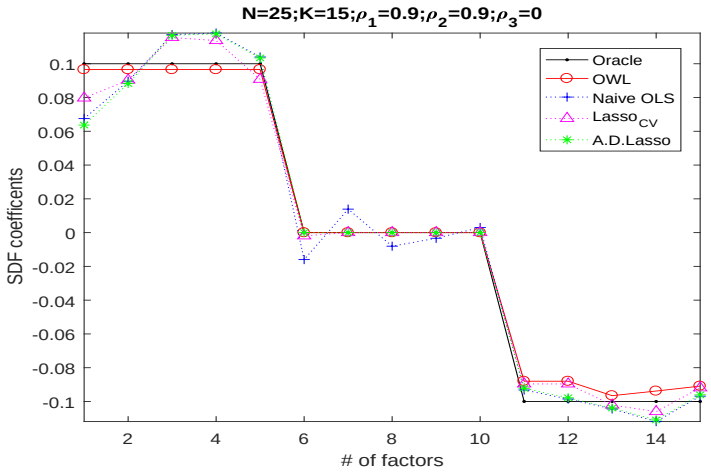
$$\rho = \begin{pmatrix} bk_1 & & 0 \\ & bk_2 & \\ 0 & & bk_3 \end{pmatrix}$$

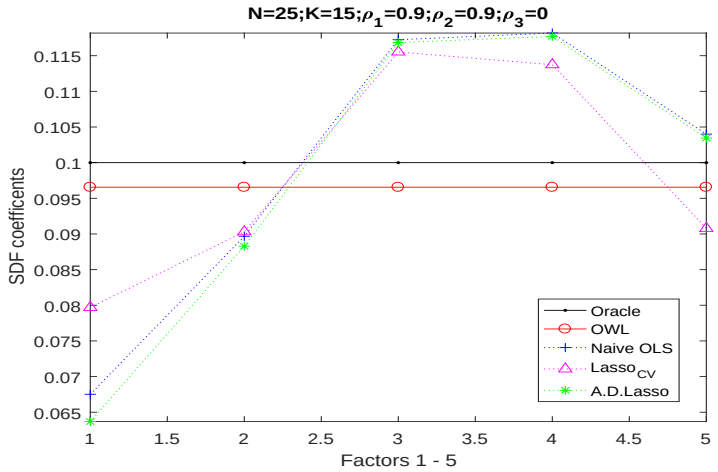
Simulation design Contd'

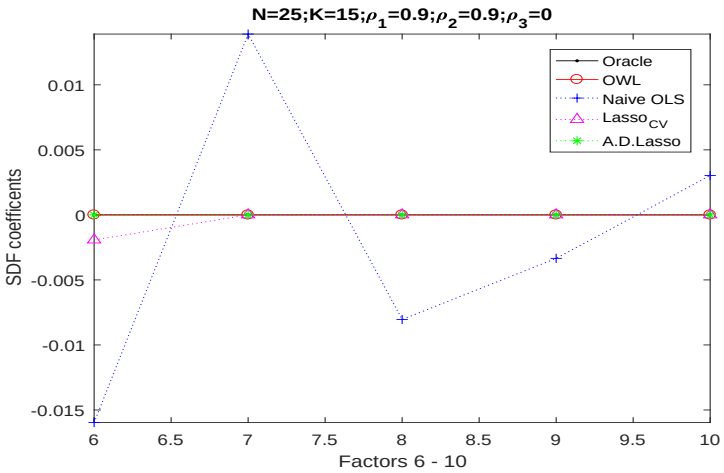
- In each block, the column vectors of C are correlated with each other, with correlation coefficient of ρ_1 , ρ_2 and ρ_3. But, these three blocks are uncorrelated with each other.
- We specify matrix ρ, and randomly generate an $N \times K$ matrix using the $i.i.d.$ Gaussian distribution. Then multiply it with the Choloski decomposition of ρ to obtain the covariance matrix C, denoted as $simC$.
- We further specify an oracle value for b, then the mean cross-section returns can be simulated by $\mu_R = simC * b + e$, where e is a $N \times 1$ $i.i.d.$ error vector with the scale about 10% of $simC$.

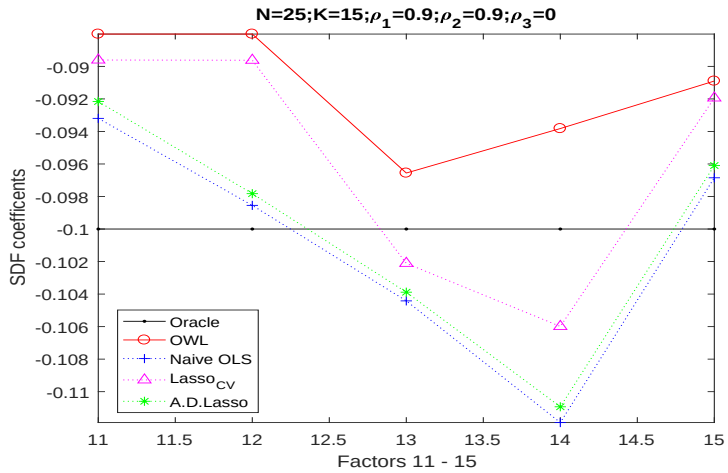
Simulation results

- Case I: $N > K$ $(N = 25, K = 15)$



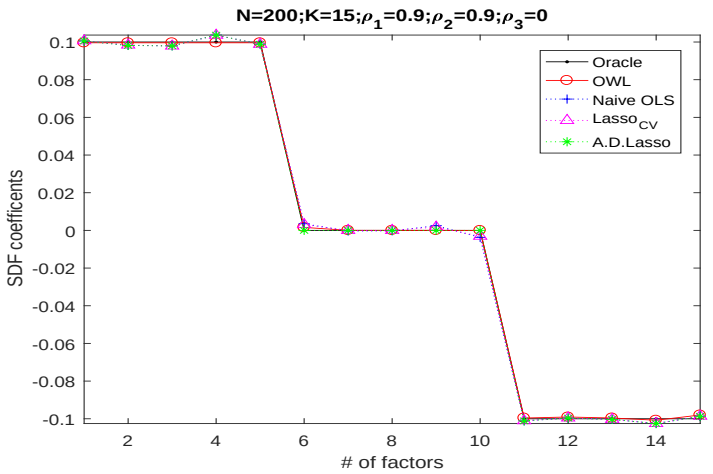


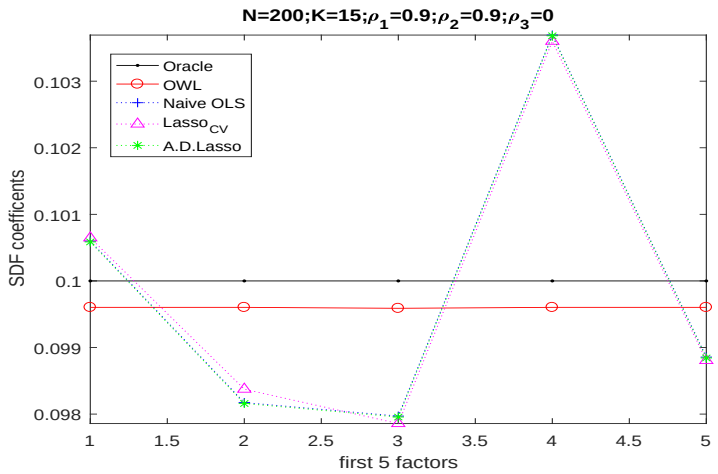


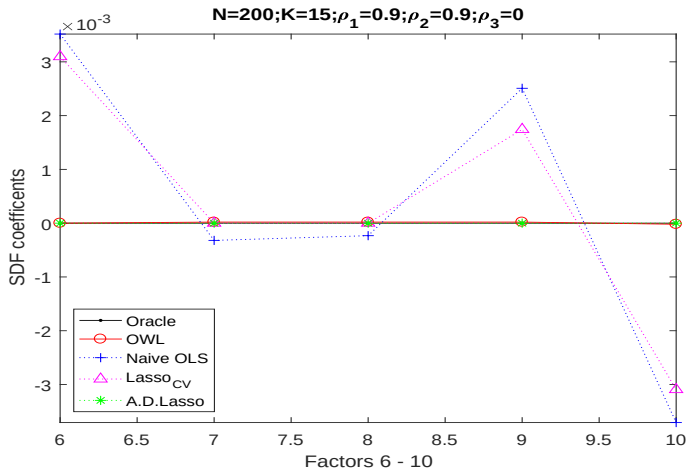


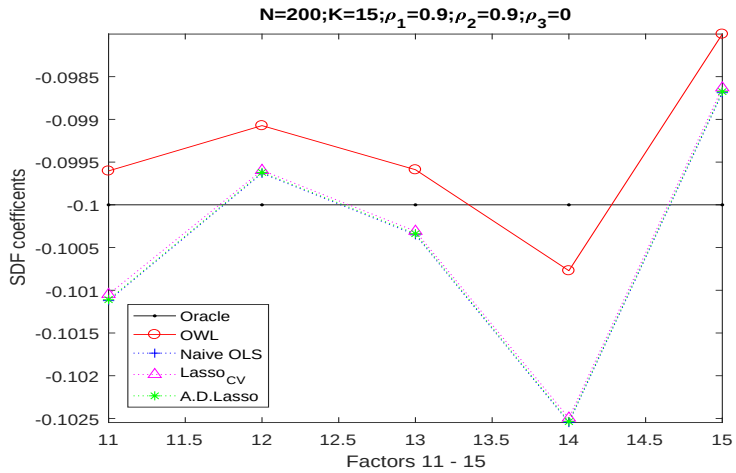
Simulation results

- Case II: $N >> K$ $(N = 200, K = 15)$









Simulation result

- OWL can successfully group highly correlated variables, and assign the same coefficients to them.
- OWL can provide satisfactory sparsity solutions, while LASSO provides inconsistent sparsity solutions.
- When K is large relative to N, OWL provides more accurate estimation than LASSO and adaptive LASSO. Adaptive LASSO depends on a consistent estimator as the adaptive weight (usually the OLS estimator), when K is large relative to N, it may be difficult to obtain a consistent estimator.

Empirical exercise

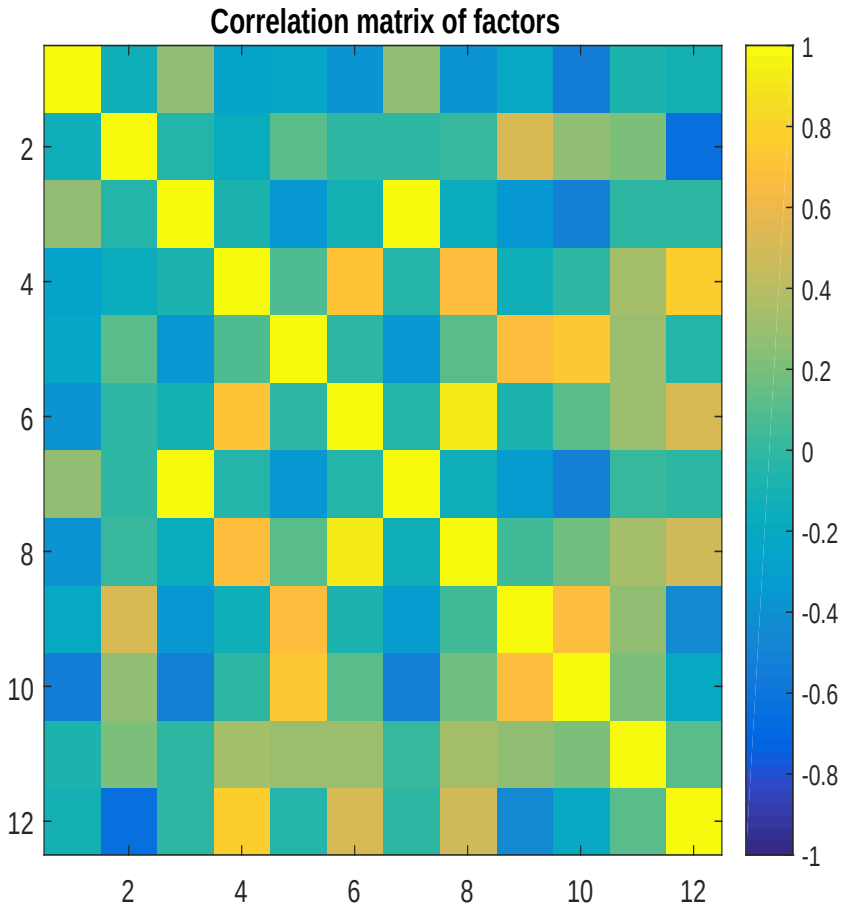
- Candidate factors: 12 popular and recently proposed factors from popular models like FF5, q-theory models etc.
- Test portfolios are bi-variate-sorted portfolios from Keneth French's online data library. Each test portfolio are sorted into 25 portfolios. We include 7 of these bi-variate sorted portfolios, plus a 49 industrial portfolio.
- The time horizon is monthly data from January 1967 to December 2016.

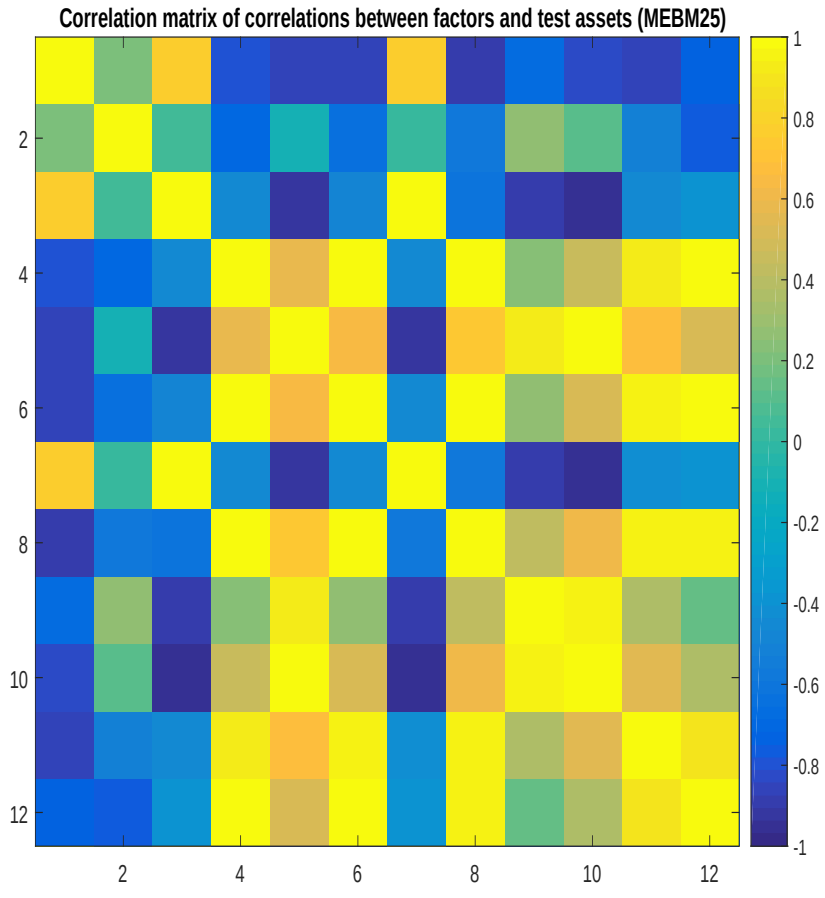
Empirical exercise

- candidate factors: {mkt, mom, smb, hml, rmw, cma, me, ia, roe, qmj, bab, hmldevil}
- 10 test portfolios: {MEBM, BMOP, MEBETA, MEINV, MENI, MEOP, MEMOM, 49INDUS, 175SORT, 175SORT-49INDUS}

Factor structure

Back





result


OWL estimators of SDF coefficients
Factors
2
4
6
8
10
12
1
2
3
4
5
6
7
8
9
10
Test portfolios
0.05
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0

comments

- The covariance matrix between factors and test assets (which matters to determine the SDF coefficients) exhibits higher dependence than the original factor correlation.
  $\Rightarrow$ half of factors exhibits correlation coefficients great than (absolute value) 0.8!
- Shrinkage result depends on test portfolio:
  $\Rightarrow$ Depending on the test portfolios, the shrinkage result is more biased to select the same characters that used to form the test portfolios.

comments contd'

- 'Mkt' is overwhelmingly the most significant factor, it was selected by all test portfolios.
- 'Momentum', 'QMJ', 'ROE' week factors: not selected in any of the 7 bi-variable sorted FF 25 portfolios.
- 'HML' is a strong factor which had been selected in many different portfolios even for those are sorted by other characters.

Comments contd'

- Fama and French's 'SMB' and Hou et al.'s 'ME' as different measures of the same (size) character, they have been grouped together by OWL, with a similar magnitude in coefficients.
- the 49 industry portfolio is not explained by any of these popular characteristics, except the 'Mkt' factor.

Next: Pooling together 7 bi-variable sorted 25 portfolios, forming a 175 test portfolio. $\rightarrow$ Two-stage testing procedure to select factors that provides independent information.

First stage: OWL estimation of 175 Bi-variable sorted portfolios

Table: OWL estimation of 175 bi-variable sorted portfolios

Candidate factors	**OWL estimator 175_sorted**
mkt	0.0321
mom	0
smb	0.0053
hml	0.02
rmw	0.0053
cma	0.002
me	0.0053
ia	0.002
roe	0
qmj	0
bab	0.0053
hmldevil	0.0053

Second stage testing using 175 Bi-variable sorted portfolios

Table: Test portfolio: 175 bi-variable sorted portfolios

	coefficient	t-stat
group 1 regression		
(Intercept)	0.1819	1.753
Mkt	0.0324	**6.29**
HML	0.0468	**10.643**
group 2 regression		
(Intercept)	-0.0767	-3.1181
SMB	0.2048	**3.4865**
RMW	0.0438	**6.3283**
ME	0.2401	**3.9173**
BAB	-0.0002	-0.0256
HMLdevil	0.0215	**4.3587**
group 3 regression		
(Intercept)	0.0191	0.8289
CMA	0.0083	0.829
IA	-0.0058	-0.3906

Conclusions