On LASSO for Predictive Regression

Jihyung Lee, Zhentao Shi and Zhan Gao

University of Illinois
Chinese University of Hong Kong
University of Southern California

Predictive regression

- Prediction is one of the fundamental tasks of econometrics.
- Predictive regression in financial markets

$$
\begin{aligned}
y_i &= \beta_1^* + x_i\beta_2^* + u_i \\
x_i &= x_{i-1} + e_i
\end{aligned}
$$

  - Unconventional inference
  - Weak signal
  - Dilemma in variable selection
- Theoretical understanding of many popular machine learning methods is working in progress.

Predictive regression

- Prediction is one of the fundamental tasks of econometrics.
- Predictive regression in financial markets

$$y_i = \beta_1^* + x_i \beta_2^* + u_i$$
$$x_i = x_{i-1} + e_i$$

  - Unconventional inference
  - Weak signal
  - Dilemma in variable selection
- Theoretical understanding of many popular machine learning methods is working in progress.

Predictive regression

- Prediction is one of the fundamental tasks of econometrics.
- Predictive regression in financial markets

$$
\begin{aligned}
y_i &= \beta_1^* + x_i\beta_2^* + u_i \\
x_i &= x_{i-1} + e_i
\end{aligned}
$$

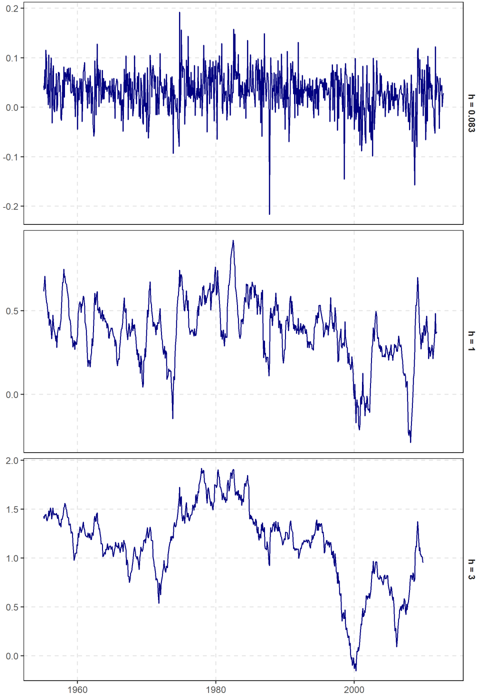
  - Unconventional inference
  - Weak signal
  - Dilemma in variable selection
- Theoretical understanding of many popular machine learning methods is working in progress.

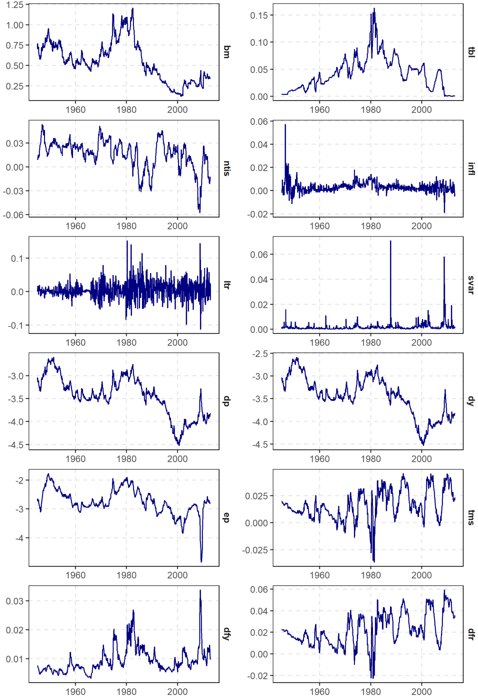
Real Data Experiment: Welch and Goyal (2008)

- Monthly data: January 1945—December 2012
- Dependent variable: S&P 500 excess return
- 12 predictors:
  - long-term return of government bonds (`ltr`), stock variance (`svar`), inflation (`infl`), dividend price ratio (`dp`), dividend yield (`dy`), earning price ratio (`ep`), term spread (`tms`), default yield spread (`dfy`), default return spread (`dfr`), book-to-market ratio (`bm`), and treasury bill rates (`tbl`).

Real Data Experiment: Welch and Goyal (2008)

- Monthly data: January 1945—December 2012
- Dependent variable: S&P 500 excess return
- 12 predictors:
  - long-term return of government bonds (`ltr`), stock variance (`svar`), inflation (`infl`), dividend price ratio (`dp`), dividend yield (`dy`), earning price ratio (`ep`), term spread (`tms`), default yield spread (`dfy`), default return spread (`dfr`), book-to-market ratio (`bm`), and treasury bill rates (`tbl`).


0.2
0.1
0.0
-0.1
-0.2
h = 0.083
0.5
0.0
h = 1
2.0
1.5
1.0
0.5
0.0
h = 3
1960
1980
2000


bm
tbl
ntis
infl
ltr
svar
dp
dy
ep
tms
dfy
dfr

LASSO Family

- Sample size n, indexed by i.
- Dependent variable y_i; regressor x_{ij}, $j = 1, \ldots, p$.
- LASSO-type estimators

$$\min_{\beta} \left\{ \|y - X\beta\|_2^2 + \lambda_n \sum_{j=1}^{p} \hat{\tau}_j |\beta_j| \right\}$$

- Plain LASSO (Plasso): $\hat{\tau}_j = 1$
- Standardized LASSO (Slasso): $\hat{\tau}_j =$ sample sd of $(x_{ij})_{i=1}^{n}$

LASSO Family

- Sample size n, indexed by i.
- Dependent variable y_i; regressor x_{ij}, $j = 1, \ldots, p$.
- LASSO-type estimators

$$\min_{\beta} \left\{ \|y - X\beta\|_2^2 + \lambda_n \sum_{j=1}^{p} \hat{\tau}_j |\beta_j| \right\}$$

- Plain LASSO (Plasso): $\hat{\tau}_j = 1$
- Standardized LASSO (Slasso): $\hat{\tau}_j =$ sample sd of $(x_{ij})_{i=1}^{n}$

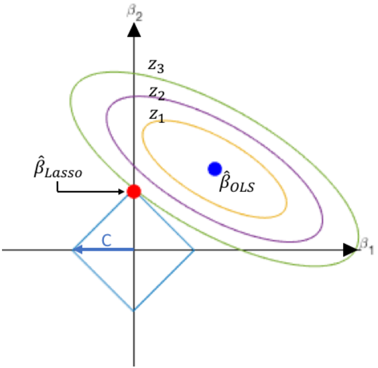
LASSO Family

- Sample size n, indexed by i.
- Dependent variable y_i; regressor x_{ij}, $j = 1, \ldots, p$.
- LASSO-type estimators

$$\min_{\beta} \left\{ \|y - X\beta\|_2^2 + \lambda_n \sum_{j=1}^{p} \hat{\tau}_j |\beta_j| \right\}$$

- Plain LASSO (Plasso): $\hat{\tau}_j = 1$
- Standardized LASSO (Slasso): $\hat{\tau}_j =$ sample sd of $(x_{ij})_{i=1}^{n}$

Variable Screening Effect



Plain LASSO is numerically equivalent to

$$\min_{\beta} \|y - X\beta\|_2^2 \quad \text{s.t.} \quad \|\beta\|_1 \le C_n$$

Oracle Property

- Active set $M^* = \{j : \beta_j^{0*} \neq 0\}$
- Inactive set $M^{*c} = \{j : \beta_j^{0*} = 0\}$
- **Oracle estimator**: Given prior knowledge about M^*, ideally

$$\widehat{\beta}^{ora} = \arg\min_{\beta} \|y - \sum_{j \in M^*} x_j \beta_j\|_2^2.$$

Oracle Property

- Active set $M^* = \{j : \beta_j^{0*} \neq 0\}$
- Inactive set $M^{*c} = \{j : \beta_j^{0*} = 0\}$
- **Oracle estimator**: Given prior knowledge about M^*, ideally

$$\widehat{\beta}^{ora} = \arg\min_{\beta} \|y - \sum_{j \in M^*} x_j \beta_j\|_2^2.$$

Adaptive LASSO

- Adaptive LASSO (Alasso) with $\hat{\tau}_j = 1/|\hat{\beta}_j^{init}|$ enjoys the oracle property.
- Alasso differentiates the penalty weight. For example if $\hat{\beta}_j^{init} = \hat{\beta}_j^{ols}$, then

$$\hat{\beta}^A = \arg\min_{\beta} \left\{ \|y - X\beta\|_2^2 + \lambda_n \sum_{j=1}^{p} \frac{1}{|\widehat{\beta}_j^{ols}|} |\beta_j| \right\}.$$

  - $\hat{\tau}_j = 1/|\widehat{\beta}_j^{ols}| = O_p(1)$ when $\beta_j^* \neq 0$;
  - $\hat{\tau}_j = 1/|\widehat{\beta}_j^{ols}| \asymp \sqrt{n}$ when $\beta_j^* = 0$

Adaptive LASSO

- Adaptive LASSO (Alasso) with $\hat{\tau}_j = 1/|\hat{\beta}_j^{init}|$ enjoys the oracle property.
- Alasso differentiates the penalty weight. For example if $\hat{\beta}_j^{init} = \hat{\beta}_j^{ols}$, then

$$\hat{\beta}^A = \arg\min_{\beta} \left\{ \|y - X\beta\|_2^2 + \lambda_n \sum_{j=1}^{p} \frac{1}{|\widehat{\beta}_j^{ols}|} |\beta_j| \right\}.$$

  - $\hat{\tau}_j = 1/|\widehat{\beta}_j^{ols}| = O_p(1)$ when $\beta_j^* \neq 0$;
  - $\hat{\tau}_j = 1/|\widehat{\beta}_j^{ols}| \asymp \sqrt{n}$ when $\beta_j^* = 0$

Contributions

- Machine learning method in new environment
- Surprisingly, Alasso does not enjoy oracle property
- Propose a simple new estimator to restore oracle property

Section 2

Pure Unit Root: An Appetizer

Simple Model: Unit Root Regressors

Linear Regression

$$y_i = \sum_{j=1}^{p} x_{ij}\beta^*_{jn} + u_i = x_{i\cdot}\beta^*_n + u_i,\ i = 1, \ldots, n$$

- For simplicity, regressors follow a pure unit root process[1]

$$x_{i\cdot} = x_{(i-1)\cdot} + e_{i\cdot} = \sum_{k=1}^{i} e_{k\cdot}.$$

[1] Paper considers the more general case of **local-to-unity**.

Simple Model: Unit Root Regressors

Linear Regression

$$y_i = \sum_{j=1}^{p} x_{ij}\beta^*_{jn} + u_i = x_{i\cdot}\beta^*_n + u_i,\ i = 1, \ldots, n$$

- For simplicity, regressors follow a pure unit root process[1]

$$x_{i\cdot} = x_{(i-1)\cdot} + e_{i\cdot} = \sum_{k=1}^{i} e_{k\cdot}.$$

[1] Paper considers the more general case of **local-to-unity**.

Balance

- True coefficient $\beta_{jn}^* = \beta_j^{0*}/\sqrt{n}$ (Paper consider the more general case $\beta_j^{0*}/n^{\delta_j}$ for $\delta_j \in (0,1)$), where β_j^{0*} is a fixed constant
  - If $\beta_j^{0*} = 0$, inactive
  - If $\beta_j^{0*} \neq 0$, active
- Asymptotic framework: p fixed and $n \to \infty$
- The OLS estimator

$$\widehat{\beta}^{ols} = \arg\min_{\beta} \|y - X\beta\|_2^2$$
$$= (X'X)^{-1} X'y$$

Balance

- True coefficient $\beta_{jn}^* = \beta_j^{0*} / \sqrt{n}$ (Paper consider the more general case $\beta_j^{0*} / n^{\delta_j}$ for $\delta_j \in (0, 1)$), where β_j^{0*} is a fixed constant
  - If $\beta_j^{0*} = 0$, inactive
  - If $\beta_j^{0*} \neq 0$, active
- Asymptotic framework: p fixed and $n \to \infty$
- The OLS estimator

$$\begin{aligned}\widehat{\beta}^{ols} &= \arg\min_{\beta} \|y - X\beta\|_2^2 \\ &= \left(X'X\right)^{-1} X'y\end{aligned}$$

Balance

- True coefficient $\beta^*_{jn} = \beta^{0*}_j / \sqrt{n}$ (Paper consider the more general case $\beta^{0*}_j / n^{\delta_j}$ for $\delta_j \in (0, 1)$), where β^{0*}_j is a fixed constant
    - If $\beta^{0*}_j = 0$, inactive
    - If $\beta^{0*}_j \neq 0$, active
- Asymptotic framework: p fixed and $n \to \infty$
- The OLS estimator

$$
\begin{aligned}
\widehat{\beta}^{ols} &= \arg\min_{\beta} \|y - X\beta\|_2^2 \\
&= \left(X'X\right)^{-1} X'y
\end{aligned}
$$

Asymptotics Distribution for OLS

The OLS limit distribution is

$$n(\hat{\beta}^{ols} - \beta_n^*) = \left(\frac{X'X}{n^2}\right)^{-1} \frac{X'u}{n} \Longrightarrow \Omega^{-1}\zeta$$

where

- $\frac{X'X}{n^2} \Longrightarrow \Omega := \int_0^1 B_x(r)B_x(r)'dr$
- $\frac{X'u}{n} \Longrightarrow \zeta, := \int_0^1 B_x(r)dB_{u^+}(r) + \int_0^1 B_x(r)\Sigma_{eu}'\Sigma_{ee}^{-1}dB_x(r)'$, $u_i^+ = u_i - \Sigma_{eu}'\Sigma_{ee}^{-1}e_{i\cdot}'$, and then $\frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor nr \rfloor} u_i^+ \Longrightarrow B_{u^+}(r)$

Asymptotics Distribution for OLS

The OLS limit distribution is

$$n(\hat{\beta}^{ols} - \beta_n^*) = \left(\frac{X'X}{n^2}\right)^{-1} \frac{X'u}{n} \Longrightarrow \Omega^{-1}\zeta$$

where

- $\frac{X'X}{n^2} \Longrightarrow \Omega := \int_0^1 B_x(r)B_x(r)'dr$
- $\frac{X'u}{n} \Longrightarrow \zeta, := \int_0^1 B_x(r)dB_{u^+}(r) + \int_0^1 B_x(r)\Sigma'_{eu}\Sigma_{ee}^{-1}dB_x(r)',$ $u_i^+ = u_i - \Sigma'_{eu}\Sigma_{ee}^{-1}e'_{i\cdot}$, and then $\frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor nr \rfloor} u_i^+ \Longrightarrow B_{u^+}(r)$

Oracle

- Oracle estimator

$$\widehat{\beta}^{ora} = \arg\min_{\beta} \|y - \sum_{j \in M^*} x_j \beta_j\|_2^2.$$

Its asymptotic distribution is

$$n\left(\hat{\beta}^{ora} - \beta^*_{M^*,n}\right) \Longrightarrow \Omega_{M^*}^{-1} \zeta_{M^*}.$$

Adaptive LASSO

Theorem

Suppose the linear model satisfies the assumption about innovations and $\widehat{\beta}_j^{init} = \widehat{\beta}_j^{ols}$. If the tuning parameter λ_n is chosen such that $\lambda_n \to \infty$ and $\lambda_n / \sqrt{n} \to 0$, then Alasso attains the ***Oracle Property:***

1. *Variable selection consistency:*

$$P(\widehat{M}^A = M^*) \to 1.$$

2. *Asymptotic distribution of $\hat{\beta}^A_{M^*}$:*

$$n(\hat{\beta}^A - \beta_n^*)_{M^*} \Longrightarrow \Omega_{M^*}^{-1}\zeta_{M^*}.$$

Section 3

Mixed Roots

Cointegration and Mixed Roots

- p_z stationary variables z_{ij}: $\mathcal{I}_0$
- p_c cointegration system x_{ij}^c: $\mathcal{C}$

$$\underset{p_1 \times p_c}{A} X_i^c = X_{1i}^c - \underset{p_1 \times p_2}{A_1} X_{2i}^c = v_{1i}$$

$$\triangle X_{2i}^c = v_{2i}$$

  - p_1 cointegration residuals v_{1ij}: $\mathcal{C}_1$
  - p_2 $(= p_c - p_1)$ unit root processes in the cointegration system x_{2ij}^c: $\mathcal{C}_2$
- p_x pure unit root processes x_{ij}: $\mathcal{I}_1$

- Asymptotic framework: $n \to \infty$ while $p = p_z + p_c + p_x$ fixed

Cointegration and Mixed Roots

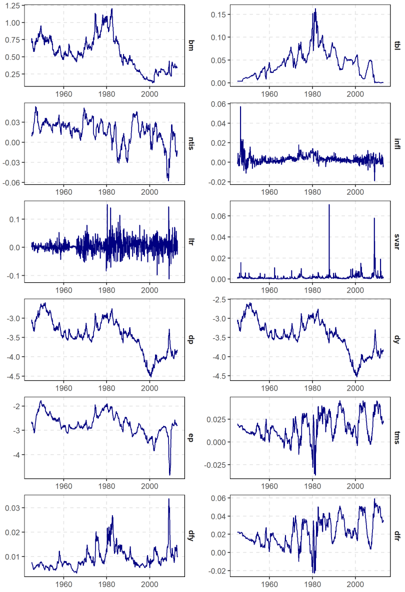
- p_z stationary variables z_{ij}: $\mathcal{I}_0$
- p_c cointegration system x_{ij}^c: $\mathcal{C}$

$$\underset{p_1 \times p_c}{A} X_i^c = X_{1i}^c - \underset{p_1 \times p_2}{A_1} X_{2i}^c = v_{1i}$$
$$\triangle X_{2i}^c = v_{2i}$$

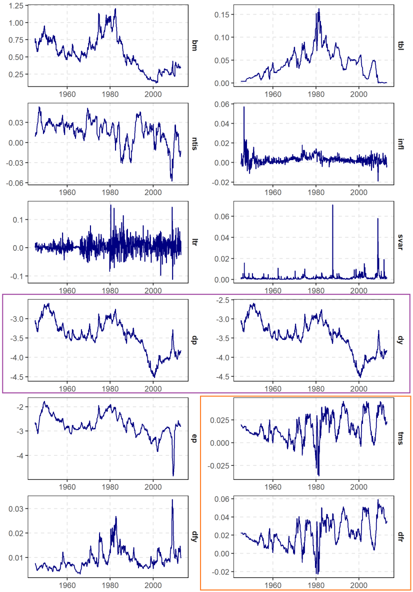
  - p_1 cointegration residuals v_{1ij}: $\mathcal{C}_1$
  - p_2 $(= p_c - p_1)$ unit root processes in the cointegration system x_{2ij}^c: $\mathcal{C}_2$
- p_x pure unit root processes x_{ij}: $\mathcal{I}_1$

- Asymptotic framework: $n \to \infty$ while $p = p_z + p_c + p_x$ fixed

Motivation of Three Types



Motivation of Three Types



Model

- Linear model (color code: green=infeasible; red=feasible)

$$
\begin{aligned}
y &= Z\alpha^* + V_1\phi_1^* + X\beta^* + u \\
&= Z\alpha^* + X_1^c\phi_1^* + X_2^c\phi_2^* + X\beta^* + u \\
&= Z\alpha^* + X^c\phi^* + X\beta^* + u \\
&= W\theta^* + u
\end{aligned}
$$

Model

- Linear model (color code: green=infeasible; red=feasible)

$$
\begin{aligned}
y &= Z\alpha^* + V_1\phi_1^* + X\beta^* + u \\
&= Z\alpha^* + X_1^c\phi_1^* + X_2^c\phi_2^* + X\beta^* + u \\
&= Z\alpha^* + X^c\phi^* + X\beta^* + u \\
&= W\theta^* + u
\end{aligned}
$$

Rotation: Bridge Between Two Worlds

- Rotation matrix $Q = \begin{pmatrix} I_{p_z} & 0 & 0 & 0 \\ 0 & I_{p_1} & 0 & 0 \\ 0 & A_1' & I_{p_2} & 0 \\ 0 & 0 & 0 & I_{p_x} \end{pmatrix}$
- $W\theta = \left(WQ^{-1}\right)(Q\theta)$
  - Rotated parameter

$$[\alpha, \phi_1, \phi_2, \beta] \overset{Q\theta}{\rightleftarrows} [\alpha, \phi_1, A_1'\phi_1 + \phi_2, \beta]$$

$$\left[\alpha^{0*}, \phi_1^{0*}, \phi_2^{0*}, \beta^{0*}/\sqrt{n}\right] \overset{Q\theta^*}{\rightleftarrows} \left[\alpha^{0*}, \phi_1^{0*}, 0, \beta^{0*}/\sqrt{n}\right]$$

  - Rotated regressor

$$[Z, X_1^c, X_2^c, X] \overset{WQ^{-1}}{\rightleftarrows} [Z, V_1, X_2^c, X]$$

OLS Theory in the Infeasible World

- "Extend" $Z^+ = (Z, V_1)$ and $X^+ = (X_2^c, X)$.
- Normalizing matrix $R_n = \mathrm{diag}\left(\sqrt{n}\mathbf{1}'_{p_z+p_1}, n\mathbf{1}'_{p_2+p_x}\right)$

Theorem

If the innovation vector follows a linear process with some regularity conditions (detailed in the paper), then

$$R_n Q(\widehat{\theta}^{ols} - \theta_n^*) = \begin{pmatrix} \sqrt{n}(\widehat{\alpha}^{ols} - \alpha^{0*}) \\ \sqrt{n}(\widehat{\phi}_1^{ols} - \phi_1^{0*}) \\ n(\widehat{\phi}_2^{ols} + A_1'\widehat{\phi}_1^{ols}) \\ n(\widehat{\beta}^{ols} - \beta_n^*) \end{pmatrix} \Longrightarrow \left(\Omega^+\right)^{-1} \zeta^+$$

OLS Theory in the Infeasible World

- "Extend" $Z^+ = (Z, V_1)$ and $X^+ = (X_2^c, X)$.
- Normalizing matrix $R_n = \text{diag}\left(\sqrt{n}\mathbf{1}'_{p_z+p_1}, n\mathbf{1}'_{p_2+p_x}\right)$

Theorem

If the innovation vector follows a linear process with some regularity conditions (detailed in the paper), then

$$R_n Q(\widehat{\theta}^{ols} - \theta_n^*) = \begin{pmatrix} \sqrt{n}(\widehat{\alpha}^{ols} - \alpha^{0*}) \\ \sqrt{n}(\widehat{\phi}_1^{ols} - \phi_1^{0*}) \\ n(\widehat{\phi}_2^{ols} + A_1' \widehat{\phi}_1^{ols}) \\ n(\widehat{\beta}^{ols} - \beta_n^*) \end{pmatrix} \Longrightarrow \left(\Omega^+\right)^{-1} \zeta^+$$

OLS Theory in the Feasible World

- In practice Q is unknown, so

$$\begin{pmatrix} \sqrt{n}(\widehat{\alpha}^{ols} - \alpha^{0*}) \\ \sqrt{n}(\widehat{\phi}_1^{ols} - \phi_1^{0*}) \\ \sqrt{n}(\widehat{\phi}_2^{ols} - \phi_2^{0*}) \\ n(\widehat{\beta}^{ols} - \beta_n^{*}) \end{pmatrix} \Longrightarrow \begin{pmatrix} I_{p_z} & 0 & 0 & 0 \\ 0 & I_{p_1} & 0 & 0 \\ 0 & -A_1' & 0 & 0 \\ 0 & 0 & 0 & I_{p_x} \end{pmatrix} \left(\Omega^{+}\right)^{-1} \zeta^{+}.$$

Index Sets in Feasible World

- $\mathcal{C} = \mathcal{C}_1 \cup \mathcal{C}_2$, $\mathcal{I} = \mathcal{I}_0 \cup \mathcal{I}_1$, and $\mathcal{M} = \{1, \ldots, p\}$
- True active set for the feasible representation $M^* = \{j : \theta_j^{0*} \neq 0\}$
- Estimated active set $\widehat{M} = \{j : \hat{\theta}_j \neq 0\}$

Index Sets in Infeasible World

- Rotation-invariant coordinates: $\mathcal{M}_Q = \mathcal{M} \backslash \mathcal{C}_2$.
    - Recall $\left[\alpha^{0*}, \phi_1^{0*}, \phi_2^{0*}, \beta^{0*}/\sqrt{n}\right] \overset{Q\theta^*}{\rightleftarrows} \left[\alpha^{0*}, \phi_1^{0*}, 0, \beta^{0*}/\sqrt{n}\right]$
- Active set:

$$M_Q^* = M^* \cap \mathcal{M}_Q$$

- Inactive set:

$$\begin{aligned} M_Q^{*c} &= \mathcal{M} \backslash M_Q^* \\ &= \mathcal{I}_0^{*c} \cup \mathcal{C}_1^{*c} \cup \mathcal{C}_2 \cup \mathcal{I}_1^{*c} \end{aligned}$$

Theorem

Suppose that the linear model satisfies Assumptions (and an extra technical assumption in the paper). Then

$$(R_n Q(\hat{\theta}^A - \theta_n^*))_{M_Q^*} \Longrightarrow (\Omega^+_{M_Q^*})^{-1} \zeta^+_{M_Q^*}$$

$$(R_n Q(\hat{\theta}^A - \theta_n^*))_{M_Q^{*c}} \xrightarrow{p} 0,$$

*and **partial** variable selection consistency*

1. $P\left(M^* \cap \mathcal{I} = \widehat{M}^A \cap \mathcal{I}\right) \to 1$
2. $P\left((M^* \cap \mathcal{C}) \subseteq (\widehat{M}^A \cap \mathcal{C})\right) \to 1,$
3. $P\left(\mathrm{CoRk}(M^*) = \mathrm{CoRk}(\widehat{M}^A)\right) \to 1.$

Overall, $M^* \subseteq \widehat{M}^A$ with high probability.

Theorem

Suppose that the linear model satisfies Assumptions (and an extra technical assumption in the paper). Then

$$(R_n Q(\hat{\theta}^A - \theta_n^*))_{M_Q^*} \Longrightarrow (\Omega^+_{M_Q^*})^{-1}\zeta^+_{M_Q^*}$$

$$(R_n Q(\hat{\theta}^A - \theta_n^*))_{M_Q^{*c}} \xrightarrow{p} 0,$$

and ***partial*** *variable selection consistency*

1. $P\left(M^* \cap \mathcal{I} = \widehat{M}^A \cap \mathcal{I}\right) \to 1$
2. $P\left((M^* \cap \mathcal{C}) \subseteq (\widehat{M}^A \cap \mathcal{C})\right) \to 1,$
3. $P\left(\mathrm{CoRk}(M^*) = \mathrm{CoRk}(\widehat{M}^A)\right) \to 1.$

Overall, $M^* \subseteq \widehat{M}^A$ with high probability.

Proof

Inspect individual Karush-Kuhn-Tucker (KKT) conditions.

- When $j \in \hat{M}^A$, the KKT condition implies

$$2W_j'(y - W\widehat{\theta}) = \mathrm{sgn}(\widehat{\theta}_j)\lambda_n \widehat{\tau}_j.$$

- LHS: j-th element of

$$2W'(y - W\widehat{\theta}) = O_p\left(Q'R_n\right) = O_p(\sqrt{n}1_{p_z}, n1_{p_c+p_x})$$

- RHS: $\mathrm{sgn}(\widehat{\theta}_j)\lambda_n / |\widehat{\theta}_j^{ols}|$.

Proof

Inspect individual Karush-Kuhn-Tucker (KKT) conditions.

- When $j \in \hat{M}^A$, the KKT condition implies

$$2W_j'(y - W\widehat{\theta}) = \text{sgn}(\widehat{\theta}_j)\lambda_n\widehat{\tau}_j.$$

  - LHS: j-th element of

$$2W'(y - W\widehat{\theta}) = O_p\left(Q'R_n\right) = O_p(\sqrt{n}1_{p_z}, n1_{p_c+p_x})$$

  - RHS: $\text{sgn}(\widehat{\theta}_j)\lambda_n / |\widehat{\theta}_j^{ols}|$.

KKT Condition

Remind $\lambda_n \to \infty$ and $\lambda_n/\sqrt{n} \to 0$. If $j \in M^{*c}$, then...

- For $j \in \mathcal{I}_0$:
    - RHS $\lambda_n/|\sqrt{n}\widehat{\theta}_j^{ols}| = \lambda_n/O_p(1) \to \infty$.
    - LHS order is $\frac{1}{\sqrt{n}} 2W_j'(y - W\widehat{\theta}) = O_p(1)$.
- For $j \in \mathcal{I}_1$:
    - RHS $\lambda_n/|n\widehat{\theta}_j^{ols}| = \lambda_n/O_p(1) \to \infty$.
    - LHS is $\frac{1}{n} 2W_j'(y - W\widehat{\theta}) = O_p(1)$.
- For $j \in \mathcal{C}$:
    - RHS $\frac{\lambda_n/\sqrt{n}}{|\sqrt{n}\widehat{\theta}_j^{ols}|} = \frac{\lambda_n/\sqrt{n}}{\text{r.v.}} \to 0$.
    - LHS $\frac{1}{n} 2W_j'(y - W\widehat{\theta})$ can degenerate.

KKT Condition

Remind $\lambda_n \to \infty$ and $\lambda_n/\sqrt{n} \to 0$. If $j \in M^{*c}$, then...

- For $j \in \mathcal{I}_0$:
  - RHS $\lambda_n / |\sqrt{n}\widehat{\theta}_j^{ols}| = \lambda_n / O_p(1) \to \infty$.
  - LHS order is $\frac{1}{\sqrt{n}} 2W_j'(y - W\widehat{\theta}) = O_p(1)$.
- For $j \in \mathcal{I}_1$:
  - RHS $\lambda_n / |n\widehat{\theta}_j^{ols}| = \lambda_n / O_p(1) \to \infty$.
  - LHS is $\frac{1}{n} 2W_j'(y - W\widehat{\theta}) = O_p(1)$.
- For $j \in \mathcal{C}$:
  - RHS $\frac{\lambda_n/\sqrt{n}}{|\sqrt{n}\widehat{\theta}_j^{ols}|} = \frac{\lambda_n/\sqrt{n}}{\text{r.v.}} \to 0$.
  - LHS $\frac{1}{n} 2W_j'(y - W\widehat{\theta})$ can degenerate.

KKT Condition

Remind $\lambda_n \to \infty$ and $\lambda_n / \sqrt{n} \to 0$. If $j \in M^{*c}$, then...

- For $j \in \mathcal{I}_0$:
  - RHS $\lambda_n / |\sqrt{n}\widehat{\theta}_j^{ols}| = \lambda_n / O_p(1) \to \infty$.
  - LHS order is $\frac{1}{\sqrt{n}} 2W_j'(y - W\widehat{\theta}) = O_p(1)$.
- For $j \in \mathcal{I}_1$:
  - RHS $\lambda_n / |n\widehat{\theta}_j^{ols}| = \lambda_n / O_p(1) \to \infty$.
  - LHS is $\frac{1}{n} 2W_j'(y - W\widehat{\theta}) = O_p(1)$.
- For $j \in \mathcal{C}$:
  - RHS $\frac{\lambda_n / \sqrt{n}}{|\sqrt{n}\widehat{\theta}_j^{ols}|} = \frac{\lambda_n / \sqrt{n}}{\text{r.v.}} \to 0$.
  - LHS $\frac{1}{n} 2W_j'(y - W\widehat{\theta})$ can degenerate.

Breaking Cointegration Link

Those $j \in C^{*c} \cap \widehat{M}^A$, mistakenly selected inactive cointegrating variables, cannot form cointegrated groups, i.e.

$$P\left(\mathrm{CoRk}(C^{*c} \cap \widehat{M}^A) = 0\right) \to 1.$$

- For simplicity, consider only one **inactive** cointegration group indexed by $\mathcal{C}$.
- Asymptotically, Alasso won't mistakenly select **all** variables in this cointegration group.

Breaking Cointegration Link

Those $j \in C^{*c} \cap \widehat{M}^A$, mistakenly selected inactive cointegrating variables, cannot form cointegrated groups, i.e.

$$P\left(\mathrm{CoRk}(C^{*c} \cap \widehat{M}^A) = 0\right) \to 1.$$

- For simplicity, consider only one **inactive** cointegration group indexed by $\mathcal{C}$.
- Asymptotically, Alasso won't mistakenly select **all** variables in this cointegration group.

Proof

Inspecting a linear combination of the corresponding KKT conditions.

- KKT condition entails $|\mathcal{C}|$ equations:

$$2x_j^{c\prime}(y - W\widehat{\theta}) = \mathrm{sgn}(\widehat{\theta}_j)\lambda_n\widehat{\tau}_j, \text{ for all } j \in \mathcal{C}.$$

- The cointegrating vector ψ linearly combines LHS of the $|\mathcal{C}|$ equations into a single equation

$$\frac{2}{\sqrt{n}}\left(\sum_{j\in\mathcal{C}} \psi_j x_j^{c\prime}\right)(y - W\widehat{\theta}) = 2\frac{v'u}{\sqrt{n}} + O_p(1)$$

- While RHS of this single equation becomes

$$\lambda_n \sum_{j\in\mathcal{C}} \frac{\mathrm{sgn}(\widehat{\theta}_j)\psi_j}{|\sqrt{n}\widehat{\theta}_j^{ols}|} = \lambda_n \sum_{j\in\mathcal{C}} \frac{\mathrm{sgn}(\widehat{\theta}_j)\psi_j}{\mathrm{r.v.}} \to \infty \text{ or } -\infty.$$

Proof

Inspecting a linear combination of the corresponding KKT conditions.

- KKT condition entails $|\mathcal{C}|$ equations:

$$2x_j^{c\prime}(y - W\widehat{\theta}) = \text{sgn}(\widehat{\theta}_j)\lambda_n\widehat{\tau}_j, \text{ for all } j \in \mathcal{C}.$$

- The cointegrating vector ψ linearly combines LHS of the $|\mathcal{C}|$ equations into a single equation

$$\frac{2}{\sqrt{n}}(\sum_{j\in\mathcal{C}} \psi_j x_j^{c\prime})(y - W\widehat{\theta}) = 2\frac{v'u}{\sqrt{n}} + O_p(1)$$

- While RHS of this single equation becomes

$$\lambda_n \sum_{j\in\mathcal{C}} \frac{\text{sgn}(\widehat{\theta}_j)\psi_j}{|\sqrt{n}\widehat{\theta}_j^{ols}|} = \lambda_n \sum_{j\in\mathcal{C}} \frac{\text{sgn}(\widehat{\theta}_j)\psi_j}{\text{r.v.}} \to \infty \text{ or } -\infty.$$

Proof

Inspecting a linear combination of the corresponding KKT conditions.

- KKT condition entails $|\mathcal{C}|$ equations:

$$2x_j^{c\prime}(y - W\widehat{\theta}) = \text{sgn}(\widehat{\theta}_j)\lambda_n\widehat{\tau}_j, \text{ for all } j \in \mathcal{C}.$$

- The cointegrating vector ψ linearly combines LHS of the $|\mathcal{C}|$ equations into a single equation

$$\frac{2}{\sqrt{n}}\Big(\sum_{j\in\mathcal{C}}\psi_j x_j^{c\prime}\Big)(y - W\widehat{\theta}) = 2\frac{v'u}{\sqrt{n}} + O_p(1)$$

- While RHS of this single equation becomes

$$\lambda_n \sum_{j\in\mathcal{C}} \frac{\text{sgn}(\widehat{\theta}_j)\psi_j}{|\sqrt{n}\widehat{\theta}_j^{ols}|} = \lambda_n \sum_{j\in\mathcal{C}} \frac{\text{sgn}(\widehat{\theta}_j)\psi_j}{\text{r.v.}} \to \infty \text{ or } -\infty.$$

Implications

- $P\left((M^* \cap \mathcal{C}) \subseteq (\widehat{M}^A \cap \mathcal{C})\right) \to 1$ is the first negative result of Alasso's oracle property
- $P\left(\mathrm{CoRk}(M^*) = \mathrm{CoRk}(\widehat{M}^A)\right) \to 1$ suggests easy remedy

Implications

- $P\left((M^* \cap \mathcal{C}) \subseteq (\widehat{M}^A \cap \mathcal{C})\right) \to 1$ is the first negative result of Alasso's oracle property
- $P\left(\mathrm{CoRk}(M^*) = \mathrm{CoRk}(\widehat{M}^A)\right) \to 1$ suggests easy remedy

Twin Adaptive LASSO (TAlasso)

- Post-selection OLS $\widehat{\theta}^{po} = \left(W'_{\widehat{M}^A} W_{\widehat{M}^A} \right)^{-1} W'_{\widehat{M}^A} y$
- Intuition: The post estimator $\widehat{\theta}_j^{po} = O_p \left(n^{-1} \right)$ for $j \in C^{*c} \cap \widehat{M}^A$
- TAlasso is a second time Alasso

$$\hat{\theta}^{TA} = \arg\min_{\theta} \left\{ \| y - W_{\widehat{M}^A} \theta \|_2^2 + \lambda_n \sum_{j \in \widehat{M}^A} \hat{\tau}_j^{po} |\theta_j| \right\}$$

where $\hat{\tau}_j^{po} = 1 / |\widehat{\theta}_j^{po}|$.

- cf: post-selection double LASSO (Belloni, Chernozhukov, and Hansen, 2014)

Twin Adaptive LASSO (TAlasso)

- Post-selection OLS $\widehat{\theta}^{po} = \left(W'_{\widehat{M}^A} W_{\widehat{M}^A}\right)^{-1} W'_{\widehat{M}^A} y$
- Intuition: The post estimator $\widehat{\theta}_j^{po} = O_p\left(n^{-1}\right)$ for $j \in C^{*c} \cap \widehat{M}^A$
- TAlasso is a second time Alasso

$$\hat{\theta}^{TA} = \arg\min_{\theta} \left\{ \|y - W_{\widehat{M}^A}\theta\|_2^2 + \lambda_n \sum_{j \in \widehat{M}^A} \hat{\tau}_j^{po} |\theta_j| \right\}$$

where $\hat{\tau}_j^{po} = 1/|\widehat{\theta}_j^{po}|$.

- cf: post-selection double LASSO (Belloni, Chernozhukov, and Hansen, 2014)

Twin Adaptive LASSO (TAlasso)

- Post-selection OLS $\widehat{\theta}^{po} = \left(W'_{\widehat{M}^A} W_{\widehat{M}^A} \right)^{-1} W'_{\widehat{M}^A} y$
- Intuition: The post estimator $\widehat{\theta}_j^{po} = O_p\left(n^{-1}\right)$ for $j \in C^{*c} \cap \widehat{M}^A$
- TAlasso is a second time Alasso

$$\hat{\theta}^{TA} = \arg\min_{\theta} \left\{ \|y - W_{\widehat{M}^A}\theta\|_2^2 + \lambda_n \sum_{j \in \widehat{M}^A} \hat{\tau}_j^{po} |\theta_j| \right\}$$

where $\hat{\tau}_j^{po} = 1/|\widehat{\theta}_j^{po}|$.

- cf: post-selection double LASSO (Belloni, Chernozhukov, and Hansen, 2014)

Oracle Property of TAlasso

Theorem

Under the same assumptions and the same rate for λ_n as in "Theorem Alasso", the TAlasso estimator $\hat{\theta}^{TA}$ satisfies

- *Asymptotic distribution:*

$$(R_n Q(\hat{\theta}^{TA} - \theta_n^*))_{M_Q^*} \Longrightarrow (\Omega^+_{M_Q^*})^{-1} \zeta^+_{M_Q^*};$$

- *Variable selection consistency:*

$$P(\widehat{M}^{TA} = M^*) \to 1.$$

Plain LASSO

Corollary

Denote $D(s, v, \beta) = s\left[v\mathrm{sgn}(\beta)I(\beta \neq 0) + |v|I(\beta = 0)\right].$

- *If* $\lambda_n/\sqrt{n} \to c_\lambda \in [0, \infty)$, *then*

$$R_n Q(\hat{\theta}^P - \theta_n^*) \Longrightarrow \arg\min_v \left\{ v'\Omega^+ v - 2v'\zeta^+ + c_\lambda \sum_{j \in \mathcal{I}_0 \cup \mathcal{C}} D(1, v_j, \theta_j^{0*}) \right\}.$$

- *If* $\lambda_n/\sqrt{n} \to \infty$ *and* $\lambda_n/n \to 0$, *then*

$$\lambda_n^{-1} R_n Q(\hat{\theta}^P - \theta_n^*) \Longrightarrow \arg\min_v \left\{ v'\Omega^+ v + \sum_{j \in \mathcal{I}_0 \cup \mathcal{C}} D(1, v_j, \theta_j^{0*}) \right\}.$$

- No way for consistent estimation and variable screening simultaneously in all three types of regressors.
- Screening occurs only on slow coefficients in $\mathcal{I}_0 \cup \mathcal{C}$.

Standardized LASSO

- For $j \in \mathcal{I}_0$, LLN gives $\widehat{\sigma}_j \xrightarrow{p} \sigma_j^0$.
- For $j \in \mathcal{C} \cup \mathcal{I}_1$, FCLT gives so that $\widehat{\sigma}_j = O_p\left(\sqrt{n}\right)$ since $\widehat{\sigma}_j / \sqrt{n} \Longrightarrow d_j := \left(\int_0^1 B_{x_j}^2(r)dr - \left(\int_0^1 B_{x_j}(r)\,dr\right)^2\right)^{1/2}$

Corollary

If $\lambda_n \to c_\lambda \in [0, \infty)$, *then*

$$R_n Q(\hat{\theta}^S - \theta_n^*) \Longrightarrow \arg\min_v \left\{ v'\Omega^+ v - 2v'\zeta^+ + c_\lambda \sum_{j \in \mathcal{C}} D(d_j, v_j, \theta_j^{0*}) \right\}.$$

If $\lambda_n \to \infty$ *and* $\lambda_n / \sqrt{n} \to 0$, *then*

$$\lambda_n^{-1} R_n Q(\hat{\theta}^S - \theta_n^*) \Longrightarrow \arg\min_v \left\{ v'\Omega^+ v + \sum_{j \in \mathcal{C}} D(d_j, v_j, \theta_j^{0*}) \right\}.$$

Section 4

Simulations

Data Generating Process

- **DGP 2 (3 types of regressors).**
- Mimic the *kitchen-sink approach* in the application.

$$y_i = \gamma^* + \sum_{l=1}^{3} z_{il}\alpha_l^* + \sum_{l=1}^{4} x_{il}^c\phi_l^* + \sum_{l=1}^{5} x_{il}\beta_{ln}^* + u_i,$$

where $\gamma^* = 0.3$, $\alpha^* = (0.4, 0, 0)$, $\phi^* = (0.5, -0.5, 0, 0)$, and $\beta_n^* = (n^{-1/2}, n^{-1/2}, 0, 0, 0)$.

Out-of-Sample MPSE

n	*Oracle*	OLS	Alas.	TAlas.	Plas.	Slas.
80	1.1479	1.3445	1.2573	**1.2497**	1.2729	1.2976
120	1.0679	1.1925	1.1346	**1.1266**	1.1523	1.1724
200	1.0350	1.1077	1.0689	**1.0651**	1.0827	1.1060
400	1.0197	1.0494	1.0389	**1.0341**	1.0444	1.0647
800	1.0162	1.0290	1.0220	**1.0193**	1.0276	1.0534

Unpredictable variance = 1

Variable Screening Success Rates

- $SR = p^{-1} \sum_{j=1}^{p} \left\{ \mathbf{1}\{\hat{\theta}_j = 0 | \theta_j^* = 0\} + \mathbf{1}\{\hat{\theta}_j \neq 0 | \theta_j^* \neq 0\} \right\}$

n	Alas.	TAlas.	Plas.	Slas.
80	0.779	**0.804**	0.643	0.572
120	0.820	**0.846**	0.634	0.584
200	0.861	**0.890**	0.617	0.593
400	0.905	**0.936**	0.593	0.601
800	0.937	**0.970**	0.576	0.606

Selection in Inactive Cointegrating Pair

For the inactive cointegrated group $\phi_3^{*0} = \phi_4^{*0} = 0$:

	$\hat{\phi}_3 = 0, \hat{\phi}_4 = 0$				$\hat{\phi}_3 \neq 0, \hat{\phi}_4 \neq 0$			
n	Alas.	TAlas.	Plas.	Slas.	Alas.	TAlas.	Plas.	Slas.
80	0.443	**0.597**	0.182	0.181	0.077	0.059	0.246	0.277
120	0.485	**0.665**	0.150	0.194	0.055	0.043	0.272	0.240
200	0.529	**0.738**	0.112	0.214	0.036	0.028	0.322	0.205
400	0.557	**0.827**	0.070	0.232	0.018	0.014	0.369	0.132
800	0.603	**0.907**	0.050	0.273	0.006	0.004	0.399	0.082

Section 5

Empirical Application

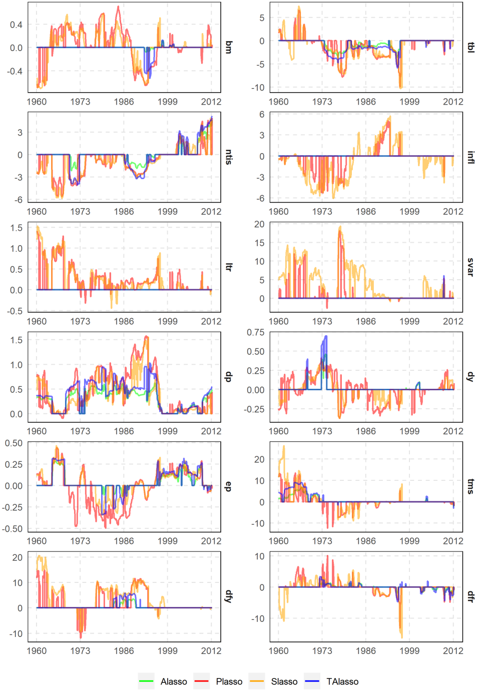
Real Data Experiment: Welch and Goyal (2008)

- Monthly data: January 1945—December 2012
- Dependent variable: S&P 500 excess return
- 12 predictors

Out-of-Sample Root MPSE

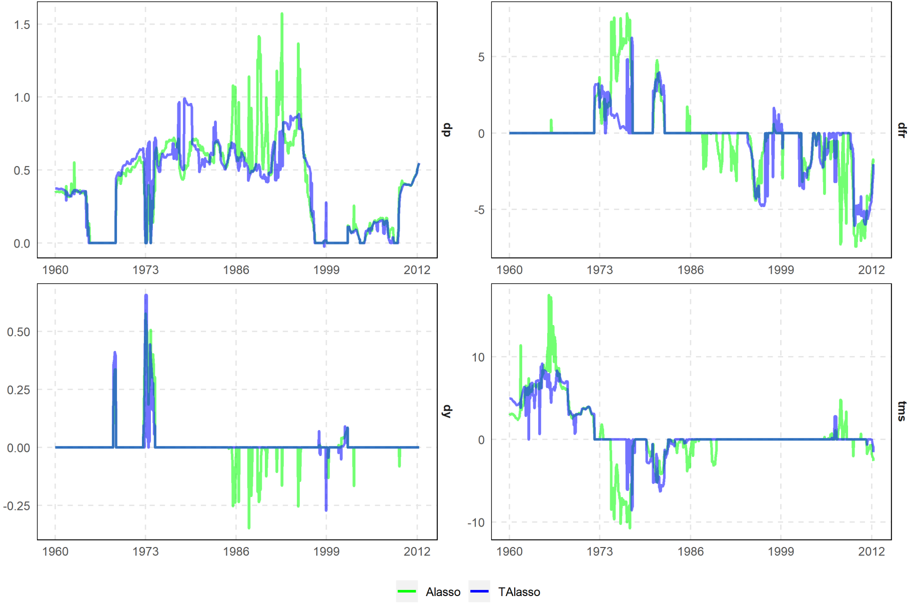
15-year rolling window. (RMPSE$\times$100)

	OLS12	OLS0	Plas.	Slas.	Alas.	TAlas.
1/12	*4.49*	4.41	4.30	4.31	**4.28**	4.33
1/4	*9.09*	8.31	8.09	8.08	**7.91**	7.94
1/2	*14.17*	13.09	13.57	13.02	**12.50**	13.00
1	19.98	*21.09*	**17.16**	19.21	18.33	20.56
2	24.38	*37.00*	22.93	23.21	20.97	**19.95**
3	33.41	*54.03*	32.53	33.47	31.82	**29.75**



bm
tbl
ntis
infl
ltr
svar
dp
dy
ep
tms
dfy
dfr
1960
1973
1986
1999
2012
Alasso
Plasso
Slasso
TAlasso

15-year Rolling Window; Horizon = 0.5



Conclusion

- Predictive regression with mixed roots
- Alasso is partially variable selection consistent
- TAlasso reclaims oracle property
- Nice finite sample performance is observed in simulations and application