On LASSO for High Dimensional Predictive Regression

Ziwei Mei and Zhentao Shi

Feb 16, 2023 @ UC Riverside

Machine Learning

- Statistics vs. econometrics
- Nonstationary time series
- Boosted Hodrick-Prescott filter
  - Phillips and Shi (2021); Mei, Phillips and Shi (2022, wp)
- LASSO in predictive regressions

$$y_t = \beta_1^* + \beta_2^* W_{t-1} + u_t$$

  - Unconventional inference with persistent W_{t-1}.
  - Lee, Shi and Gao (2022): Variable selection
  - Mei and Shi (2022, this paper): high dimension

Machine Learning

- Statistics vs. econometrics
- Nonstationary time series
- Boosted Hodrick-Prescott filter
  - Phillips and Shi (2021); Mei, Phillips and Shi (2022, wp)
- LASSO in predictive regressions

$$y_t = \beta_1^* + \beta_2^* W_{t-1} + u_t$$

  - Unconventional inference with persistent W_{t-1}.
  - Lee, Shi and Gao (2022): Variable selection
  - Mei and Shi (2022, this paper): high dimension

Machine Learning

- Statistics vs. econometrics
- Nonstationary time series
- Boosted Hodrick-Prescott filter
  - Phillips and Shi (2021); Mei, Phillips and Shi (2022, wp)
- LASSO in predictive regressions

$$y_t = \beta_1^* + \beta_2^* W_{t-1} + u_t$$

  - Unconventional inference with persistent W_{t-1}.
  - Lee, Shi and Gao (2022): Variable selection
  - Mei and Shi (2022, this paper): high dimension

Machine Learning

- Statistics vs. econometrics
- Nonstationary time series
- Boosted Hodrick-Prescott filter
  - Phillips and Shi (2021); Mei, Phillips and Shi (2022, wp)
- LASSO in predictive regressions

$$y_t = \beta_1^* + \beta_2^* W_{t-1} + u_t$$

  - Unconventional inference with persistent W_{t-1}.
  - Lee, Shi and Gao (2022): Variable selection
  - Mei and Shi (2022, this paper): high dimension

Machine Learning

- Statistics vs. econometrics
- Nonstationary time series
- Boosted Hodrick-Prescott filter
    - Phillips and Shi (2021); Mei, Phillips and Shi (2022, wp)
- LASSO in predictive regressions

$$y_t = \beta_1^* + \beta_2^* W_{t-1} + u_t$$

- Unconventional inference with persistent W_{t-1}.
- Lee, Shi and Gao (2022): Variable selection
- Mei and Shi (2022, this paper): high dimension

Real Data Examples

- Finance
  - Welch and Goyal (2008); used in Lee, Shi and Gao (2022)
  - Dependent variable: S&P 500 excess return
  - 12 predictors
- Macroeconomics
  - Medeiros, Vasoncelos, Veiga, and Zilberman (2021)
  - FRED-MD database
  - Dependent variable: Inflation (CPI)
  - About 500 constructed predictors

Real Data Examples

- Finance
  - Welch and Goyal (2008); used in Lee, Shi and Gao (2022)
  - Dependent variable: S&P 500 excess return
  - 12 predictors
- Macroeconomics
  - Medeiros, Vasoncelos, Veiga, and Zilberman (2021)
  - FRED-MD database
  - Dependent variable: Inflation (CPI)
  - About 500 constructed predictors

LASSO Family

- Sample size n, indexed by t.
- Regressor W_{jt}, $j = 1, \ldots, p$.
- Plain LASSO (Plasso).

$$(\widehat{\alpha}, \widehat{\theta}) = \arg\min_{\alpha,\theta} \left\{ n^{-1} \left\| Y - \alpha 1_n - W\theta \right\|_2^2 + \lambda \left\| \theta \right\|_1 \right\},$$

Prediction is made as

$$\widehat{y}_{n+1} = \widehat{\alpha} + W_n^\top \widehat{\theta}$$

- Undesirable property: Estimate varies with scale of W_{jt}.

LASSO Family

- Sample size n, indexed by t.
- Regressor W_{jt}, $j = 1, \dots, p$.
- Plain LASSO (Plasso).

$$(\widehat{\alpha}, \widehat{\theta}) = \arg\min_{\alpha, \theta} \left\{ n^{-1} \left\| Y - \alpha 1_n - W\theta \right\|_2^2 + \lambda \left\| \theta \right\|_1 \right\},$$

Prediction is made as

$$\widehat{y}_{n+1} = \widehat{\alpha} + W_n^\top \widehat{\theta}$$

- Undesirable property: Estimate varies with scale of W_{jt}.

Standardized LASSO

- Default option in most statistical software
- Sample s.d. $\widehat{\sigma}_j$
- Transform W_{jt} into $W_{jt}/\widehat{\sigma}_j$
- Let $\tilde{W} = WD^{-1}$ where $D = \mathrm{diag}(\widehat{\sigma}_1, \widehat{\sigma}_2, \cdots, \widehat{\sigma}_p)$
- Standardized LASSO (Slasso)

$$(\tilde{\alpha}, \tilde{\theta}) = \arg\min_{\alpha,\theta} \left\{ n^{-1} \left\| Y - \alpha 1_n - \tilde{W}\theta \right\|_2^2 + \lambda \left\| \theta \right\|_1 \right\}$$

and makes prediction

$$\tilde{y}_{n+1} = \tilde{\alpha} + \tilde{W}_n^{\top} \tilde{\theta}$$

Standardized LASSO

- Default option in most statistical software
- Sample s.d. $\widehat{\sigma}_j$
- Transform W_{jt} into $W_{jt}/\widehat{\sigma}_j$
- Let $\tilde{W} = WD^{-1}$ where $D = \mathrm{diag}(\widehat{\sigma}_1, \widehat{\sigma}_2, \cdots, \widehat{\sigma}_p)$
- Standardized LASSO (Slasso)

$$(\tilde{\alpha}, \tilde{\theta}) = \arg\min_{\alpha,\theta} \left\{ n^{-1} \left\| Y - \alpha 1_n - \tilde{W}\theta \right\|_2^2 + \lambda \left\| \theta \right\|_1 \right\}$$

and makes prediction

$$\tilde{y}_{n+1} = \tilde{\alpha} + \tilde{W}_n^\top \tilde{\theta}$$

Section 2

Setup

True Coefficients

- DGP with true parameters (α^*, θ^*):

$$Y_t = \alpha^* + W_{t-1}^\top \theta^* + u_t$$

- Estimators
  - Plasso: $\widehat{\theta}$ estimates the original parameter θ^*
  - Slasso: $\tilde{\theta}$ estimates the transformed parameter $\tilde{\theta}^* = D\theta^*$.
- We will consider W being a mixture of I(1) and I(0), as in the application.
- But let us start with unit root regressors only...

True Coefficients

- DGP with true parameters (α^*, θ^*):

$$Y_t = \alpha^* + W_{t-1}^\top \theta^* + u_t$$

- Estimators
  - Plasso: $\widehat{\theta}$ estimates the original parameter θ^*
  - Slasso: $\tilde{\theta}$ estimates the transformed parameter $\tilde{\theta}^* = D\theta^*$.
- We will consider W being a mixture of I(1) and I(0), as in the application.
- But let us start with unit root regressors only...

OLS

- Pure unit roots (Ignore the intercept for simplicity)

$$Y_t = X_{t-1}^\top \beta^* + u_t$$

where X_{t-1} is a vector of **unit root processes**

- OLS with fixed p

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

where

- $\frac{X^\top X}{n^2} \Longrightarrow \Omega := \int_0^1 B_x(r) B_x(r)^\top dr$ (Gram matrix)
- $\frac{X^\top u}{n} \Longrightarrow \zeta := \int_0^1 B_x(r) dB_{u^+}(r) + bias$ (Empirical process)

High Dimension

- High dimensionality allows $p > n$
- Sparsity index: s is the number of non-zero coefficients $\theta_j^* \neq 0$
- Gram matrix: the sample covariance matrix $\check{\Sigma} = \check{W}^\top \check{W}/n$
  - Restriction is needed as $\check{\Sigma}$ rank deficient when $p > n$

Two Building Blocks

Definition (RE)

Restricted eigenvalue: (Bickel, Ritov and Tsybakov, 2009)

$$\kappa(\check{\Sigma}, s) = \inf_{\delta \in \mathcal{R}(s)} \frac{\delta^\top \check{\Sigma} \delta}{\delta^\top \delta}$$

where $\mathcal{R}(s) = \{\delta \in \mathbb{R}^p : \|\delta_{\mathcal{M}^c}\|_1 \leq 3\|\delta_{\mathcal{M}}\|_1,$ for all $|\mathcal{M}| \leq s\}$.

Definition (DB)

Deviation bound: An upper bound of $\|n^{-1} \sum_{t=1}^{n} \check{W}_{t-1} u_t\|_\infty$.

Two Building Blocks

Definition (RE)

Restricted eigenvalue: (Bickel, Ritov and Tsybakov, 2009)

$$\kappa(\check{\Sigma}, s) = \inf_{\delta \in \mathcal{R}(s)} \frac{\delta^\top \check{\Sigma} \delta}{\delta^\top \delta}$$

where $\mathcal{R}(s) = \{\delta \in \mathbb{R}^p : \|\delta_{\mathcal{M}^c}\|_1 \leq 3\|\delta_{\mathcal{M}}\|_1,$ for all $|\mathcal{M}| \leq s\}$.

Definition (DB)

Deviation bound: An upper bound of $\|n^{-1} \sum_{t=1}^{n} \check{W}_{t-1} u_t\|_\infty$.

Finite Sample Result

Lemma (Bühlmann and van der Geer, 2011)

If $\lambda \geq 4\|n^{-1}\sum_{t=1}^{n}\check{W}_{t-1}u_t\|_\infty$, *then*

$$
\begin{aligned}
n^{-1}\|\check{W}(\check{\theta}-\check{\theta}^*)\|_2^2 &\leq 4\lambda^2 s/\check{\kappa} \\
\|\check{\theta}-\check{\theta}^*\|_1 &\leq 4\lambda s/\check{\kappa} \\
\|\check{\theta}-\check{\theta}^*\|_2 &\leq 2\lambda\sqrt{s}/\check{\kappa},
\end{aligned}
$$

where $\check{\kappa} = \kappa(\check{\Sigma}, s)$.

- Convergence rate depends on λ, s and $\check{\kappa}$.

Literature

LASSO in predictive regression

- Koo, Anderson, Seo, and Yao (2020)
- Lee, Shi and Gao (2022)
- Fan, Lee, and Shin (2023)
- **Wijler (2022, wp)**
  - Pure unit roots regressors $X_t = X_{t-1} + e_t$
  - Innovation $e_{jt} \sim iid\, \mathcal{N}(0, \sigma^2)$ across j and t.

Contributions

- In a unified framework:
  - Analyze both Plasso and Slasso
  - Unit root regressors

$$Y_t = \alpha + X_{t-1}^\top \beta^* + u_t$$

and mixed I(1) and I(0)

$$Y_t = \alpha^* + X_{t-1}^\top \beta^* + Z_{t-1}^\top \gamma^* + u_t.$$

- A new RE
  - Based on random matrix theory
  - Low level conditions
  - Non-Gaussian, time dependent innovations
- Empirical application highlights importance of domain expertise

Contributions

- In a unified framework:
  - Analyze both Plasso and Slasso
  - Unit root regressors

$$Y_t = \alpha + X_{t-1}^\top \beta^* + u_t$$

  and mixed I(1) and I(0)

$$Y_t = \alpha^* + X_{t-1}^\top \beta^* + Z_{t-1}^\top \gamma^* + u_t.$$

- A new RE
  - Based on random matrix theory
  - Low level conditions
  - Non-Gaussian, time dependent innovations
- Empirical application highlights importance of domain expertise

Contributions

- In a unified framework:
  - Analyze both Plasso and Slasso
  - Unit root regressors

$$Y_t = \alpha + X_{t-1}^{\top}\beta^* + u_t$$

  and mixed I(1) and I(0)

$$Y_t = \alpha^* + X_{t-1}^{\top}\beta^* + Z_{t-1}^{\top}\gamma^* + u_t.$$

- A new RE
  - Based on random matrix theory
  - Low level conditions
  - Non-Gaussian, time dependent innovations
- Empirical application highlights importance of domain expertise

Section 3

Theory

From Innovation to Unit Root

$$\underset{(n\times p)}{X} = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ 1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 1 & \cdots & 1 & 0 \\ 1 & 1 & \cdots & 1 & 1 \end{pmatrix} \underset{(n\times p)}{e} = \underset{(n\times n)}{R}\, e$$

- Consider the special case $\underset{(p\times 1)}{e_t} \sim iid\, \mathcal{N}(0, \Omega)$

Unit Root Transformation

- For simplicity, ignore intercept and set $\Omega = I$.
- RE $\widehat{\kappa} = \kappa(\widehat{\Sigma}, s)$, where $\widehat{\Sigma} = X^\top X/n$.

$$
\begin{aligned}
\delta^\top \widehat{\Sigma} \delta &= n^{-1} \delta^\top \left[e^\top R^\top R e \right] \delta \\
&= n^{-1} \delta^\top [e^\top V] \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n) [V^\top e] \delta \\
&\geq n^{-1} \delta^\top [e^\top V_{1:\ell}] \mathrm{diag}(\lambda_1, \cdots, \lambda_\ell) [V_{1:\ell}^\top e] \delta \\
&\geq n^{-1} \lambda_\ell \cdot \delta^\top e^\top V_{1:\ell} V_{1:\ell}^\top e \delta \\
&\sim \left[\lambda_\ell \frac{\ell}{n} \right] \cdot \delta^\top \frac{\mathrm{Wishart}_p (I, \ell)}{\ell} \delta
\end{aligned}
$$

where $\underset{(n\times\ell)}{V_{1:\ell}}$ is the first ℓ columns of V, which reduces $\underset{(n\times p)}{V^\top e}$ to $\underset{(\ell\times p)}{V_{1:\ell}^\top e}$ of iid $\mathcal{N}(0,1)$ entries.

Order of RE

Fact (Smeekes and Wijler, 2021)

The ℓth eigenvalue of $R^\top R$ is

$$\lambda_\ell = 0.5 / \left[1 - \cos\left(\frac{(2\ell - 1)\pi}{2n+1}\right)\right] \asymp \frac{n^2}{\ell^2}.$$

As a result, $\lambda_\ell \frac{\ell}{n} \asymp \frac{n}{\ell}$.

- Next, consider restricted sparse δ such that $\|\delta\|_0 \leq 2s$:

$$\delta^\top \left[\text{Wishart}_p(I, \ell) / \ell\right] \delta \overset{\text{spar.}}{\sim} \delta^\top \left[\text{Wishart}_{2s}(I, \ell) / \ell\right] \delta$$

Order of RE

Fact (Smeekes and Wijler, 2021)

The ℓth eigenvalue of $R^\top R$ is

$$\lambda_\ell = 0.5 / \left[1 - \cos\left(\frac{(2\ell - 1)\pi}{2n+1}\right)\right] \asymp \frac{n^2}{\ell^2}.$$

As a result, $\lambda_\ell \frac{\ell}{n} \asymp \frac{n}{\ell}$.

- Next, consider restricted sparse δ such that $\|\delta\|_0 \leq 2s$:

$$\delta^\top \left[\text{Wishart}_p\left(I, \ell\right) / \ell\right] \delta \overset{\text{spar.}}{\sim} \delta^\top \left[\text{Wishart}_{2s}\left(I, \ell\right) / \ell\right] \delta$$

Key Tool: Random Matrix Theory

- If a generic $\underset{(q\times 1)}{z_s} \sim iid\ \mathcal{N}(0,\Omega)$ for some full rank Ω, then

$$\widehat{\Omega} = \ell^{-1}\sum_{s=1}^{\ell} z_s z_s^{\top} \sim \text{Wishart}_q(\Omega,\ell)/\ell$$

Fact (Wainright, 2019)

When $\ell > q$, for all $c \in (0,1)$:

$$\Pr\left(\lambda_{\min}^{1/2}(\widehat{\Omega}) \leq \lambda_{\min}^{1/2}(\Omega)(1-c) - \sqrt{\ell^{-1}\cdot \mathrm{tr}(\Omega)}\right) \leq e^{-\ell c^2/2}$$

- For all $2s$-submatrices, there are $K = \binom{p}{2s} \leq p^{2s}$ choices
- Uniform bound for minimum eigenvalues of all $2s$-submatrices

Key Tool: Random Matrix Theory

- If a generic $\underset{(q\times 1)}{z_s} \sim iid\ \mathcal{N}(0,\Omega)$ for some full rank Ω, then

$$\widehat{\Omega} = \ell^{-1}\sum_{s=1}^{\ell} z_s z_s^\top \sim \text{Wishart}_q(\Omega,\ell)/\ell$$

Fact (Wainright, 2019)

When $\ell > q$, for all $c \in (0,1)$:

$$\Pr\left(\lambda_{\min}^{1/2}(\widehat{\Omega}) \le \lambda_{\min}^{1/2}(\Omega)(1-c) - \sqrt{\ell^{-1}\cdot \mathrm{tr}(\Omega)}\right) \le e^{-\ell c^2/2}$$

- For all $2s$-submatrices, there are $K = \binom{p}{2s} \le p^{2s}$ choices
- Uniform bound for minimum eigenvalues of all $2s$-submatrices

Key Tool: Random Matrix Theory

- If a generic $\underset{(q\times 1)}{z_s} \sim iid\ \mathcal{N}(0,\Omega)$ for some full rank Ω, then

$$\widehat{\Omega} = \ell^{-1}\sum_{s=1}^{\ell} z_s z_s^\top \sim \text{Wishart}_q(\Omega,\ell)/\ell$$

Fact (Wainright, 2019)

When $\ell > q$, for all $c \in (0,1)$:

$$\Pr\left(\lambda_{\min}^{1/2}(\widehat{\Omega}) \le \lambda_{\min}^{1/2}(\Omega)(1-c) - \sqrt{\ell^{-1}\cdot \mathrm{tr}(\Omega)}\right) \le e^{-\ell c^2/2}$$

- For all $2s$-submatrices, there are $K = \binom{p}{2s} \le p^{2s}$ choices
- Uniform bound for minimum eigenvalues of all $2s$-submatrices

- Demonstration with $\Omega = I$.
- Set $c = 0.5$, $\ell = 32s\log p$
- $\lambda_{\min}(\widehat{\Omega}_k) \geq 0.16$ uniformly for all $2s$-submatrices w.p.a.1.

$$
\begin{aligned}
&\Pr\left(\cup_{k\leq K}\left\{\lambda_{\min}^{1/2}(\widehat{\Omega}_k) \leq 0.4\right\}\right) \\
\leq &\Pr\left(\cup_{k\leq K}\left\{\lambda_{\min}^{1/2}(\widehat{\Omega}_k) \leq 0.5 - \sqrt{2s/\ell}\right\}\right) \\
\leq &\sum_{k\leq K}\Pr\left(\lambda_{\min}^{1/2}(\widehat{\Omega}_k) \leq 0.5 - \sqrt{2s/\ell}\right) \\
\leq & p^{2s} \times \mathrm{e}^{-32s\log p\cdot 0.5^2/2} = p^{-2s} \to 0.
\end{aligned}
$$

As a result, w.p.a.1

$$
\frac{\delta^\top\widehat{\Sigma}\delta}{\delta^\top\delta} \geq \frac{n}{\delta^\top\delta\ell}\cdot\delta^\top\frac{\mathrm{Wishart}_{2s}(I_{2s},\ell)}{\ell}\delta \geq \frac{0.16\times n}{32s\log p} = \frac{0.005n}{s\log p}
$$

- Demonstration with $\Omega = I$.
- Set $c = 0.5$, $\ell = 32 s \log p$
- $\lambda_{\min}(\widehat{\Omega}_k) \geq 0.16$ uniformly for all $2s$-submatrices w.p.a.1.

$$
\begin{aligned}
& \Pr\left(\cup_{k \leq K} \left\{ \lambda_{\min}^{1/2}(\widehat{\Omega}_k) \leq 0.4 \right\} \right) \\
\leq & \Pr\left(\cup_{k \leq K} \left\{ \lambda_{\min}^{1/2}(\widehat{\Omega}_k) \leq 0.5 - \sqrt{2s/\ell} \right\} \right) \\
\leq & \sum_{k \leq K} \Pr\left(\lambda_{\min}^{1/2}(\widehat{\Omega}_k) \leq 0.5 - \sqrt{2s/\ell} \right) \\
\leq & p^{2s} \times e^{-32 s \log p \cdot 0.5^2/2} = p^{-2s} \to 0.
\end{aligned}
$$

As a result, w.p.a.1

$$
\frac{\delta^\top \widehat{\Sigma} \delta}{\delta^\top \delta} \geq \frac{n}{\delta^\top \delta \ell} \cdot \delta^\top \frac{\text{Wishart}_{2s}(I_{2s}, \ell)}{\ell} \delta \geq \frac{0.16 \times n}{32 s \log p} = \frac{0.005 n}{s \log p}
$$

RE Bound under Gaussian Shocks

- Asymptotic framework: $n \to \infty$, and $s, p \to \infty$.
- Innovation

$$\begin{pmatrix} e_t \\ u_t \end{pmatrix} = \underset{(p+1)\times(p+1)}{\Phi} \underset{(p+1)}{\varepsilon_t}$$

- **Assumption** (Cross sectional dependence): All singular values of Φ bounded away from 0 and ∞. ($\Omega = \Phi_{1:p}\Phi_{1:p}^\top$)

Proposition

If $\varepsilon_{jt} \sim$ iid $\mathcal{N}(0,1)$ and $s/(p \wedge n) \to 0$, then under Assumption as $n \to \infty$:

$$\widehat{\kappa}/n \overset{\text{p}}{\succcurlyeq} (s \log p)^{-1}.$$

RE Bound under Gaussian Shocks

- Asymptotic framework: $n \to \infty$, and $s, p \to \infty$.
- Innovation
$$\begin{pmatrix} e_t \\ u_t \end{pmatrix} = \underset{(p+1)\times(p+1)}{\Phi} \underset{(p+1)}{\varepsilon_t}$$
- **Assumption** (Cross sectional dependence): All singular values of Φ bounded away from 0 and ∞. ($\Omega = \Phi_{1:p}\Phi_{1:p}^{\top}$)

Proposition

If $\varepsilon_{jt} \sim$ *iid* $\mathcal{N}(0,1)$ *and* $s/(p \wedge n) \to 0$, *then under Assumption as* $n \to \infty$*:*
$$\widehat{\kappa}/n \overset{\mathrm{p}}{\succcurlyeq} (s \log p)^{-1}.$$

Gaussian Approximation

- Linear process $\varepsilon_{jt} = \sum_{d=0}^{\infty} \psi_{jd}\eta_{j,t-d}$ for all j, with iid shocks (η_{jt}) is iid.
- Functional central limit theorem:

$$\frac{1}{\sqrt{n}} \sum_{s=0}^{\lfloor n\cdot \rfloor} \varepsilon_{js} \Longrightarrow \psi_j(1)B_j(\cdot)$$

- Skorokhod's representation theorem.
- RE under normal distribution carries over to non-Gaussian case if approximation holds uniformly over all j.

Gaussian Approximation

- Linear process $\varepsilon_{jt} = \sum_{d=0}^{\infty} \psi_{jd}\eta_{j,t-d}$ for all j, with iid shocks (η_{jt}) is iid.
- Functional central limit theorem:

$$\frac{1}{\sqrt{n}} \sum_{s=0}^{\lfloor n\cdot \rfloor} \varepsilon_{js} \Longrightarrow \psi_j(1) B_j(\cdot)$$

- Skorokhod's representation theorem.
- RE under normal distribution carries over to non-Gaussian case if approximation holds uniformly over all j.

Assumptions

All symbols of $c_{\sup}$ are absolute constants.

- **Assumption** (sub-exponential tail): for all j.

$$\Pr(|\eta_{jt}| > \mu) \leq C_\eta \exp\left[-\mu / c_\eta\right].$$

- **Assumption** (temporal dependence): There is $r > 0$ such that for all j

$$|\psi_{jd}| \leq C_\psi \exp\left(-c_\psi d^r\right), \quad d \in \mathbb{N}.$$

- **Assumption** (Size of model): $s \to \infty$ and $s^9/n \to 0$, and $p = O\left(n^v\right)$ for some $v \in (0, \infty)$.
- Done with RE. Move on DB.

Assumptions

All symbols of $c_{\sup}$ are absolute constants.

- **Assumption** (sub-exponential tail): for all j.

$$\Pr(|\eta_{jt}| > \mu) \leq C_\eta \exp\left[-\mu/c_\eta\right].$$

- **Assumption** (temporal dependence): There is $r > 0$ such that for all j

$$|\psi_{jd}| \leq C_\psi \exp\left(-c_\psi d^r\right), \quad d \in \mathbb{N}.$$

- **Assumption** (Size of model): $s \to \infty$ and $s^9/n \to 0$, and $p = O(n^\nu)$ for some $\nu \in (0, \infty)$.
- Done with RE. Move on DB.

Assumptions

All symbols of $c_{\sup}$ are absolute constants.

- **Assumption** (sub-exponential tail): for all j.

$$\Pr(|\eta_{jt}| > \mu) \leq C_\eta \exp\left[-\mu / c_\eta\right].$$

- **Assumption** (temporal dependence): There is $r > 0$ such that for all j

$$|\psi_{jd}| \leq C_\psi \exp\left(-c_\psi d^r\right), \quad d \in \mathbb{N}.$$

- **Assumption** (Size of model): $s \to \infty$ and $s^9/n \to 0$, and $p = O\left(n^\nu\right)$ for some $\nu \in (0, \infty)$.
- Done with RE. Move on DB.

Assumptions

All symbols of $c_{\sup}$ are absolute constants.

- **Assumption** (sub-exponential tail): for all j.

$$\Pr(|\eta_{jt}| > \mu) \leq C_\eta \exp\left[-\mu/c_\eta\right].$$

- **Assumption** (temporal dependence): There is $r > 0$ such that for all j

$$|\psi_{jd}| \leq C_\psi \exp\left(-c_\psi d^r\right), \quad d \in \mathbb{N}.$$

- **Assumption** (Size of model): $s \to \infty$ and $s^9/n \to 0$, and $p = O\left(n^\nu\right)$ for some $\nu \in (0, \infty)$.
- Done with RE. Move on DB.

Deviation Bound

Proposition

Under above Assumptions:

$$\left\| \frac{1}{n} \sum_{t=1}^{n} X_{t-1} u_t \right\|_\infty \leq C_{\text{DB}} (\log p)^{1+\frac{1}{2r}}$$

w.p.a.1.

- RE and DB ready. Move on to Plasso.

Deviation Bound

Proposition

Under above Assumptions:

$$\left\| \frac{1}{n} \sum_{t=1}^{n} X_{t-1} u_t \right\|_\infty \leq C_{\mathrm{DB}} (\log p)^{1+\frac{1}{2r}}$$

w.p.a.1.

- RE and DB ready. Move on to Plasso.

Plasso for unit roots

Theorem

If we choose $\lambda = C_{\mathrm{DB}}(\log p)^{1+\frac{1}{2r}}$ *the Plasso estimator satisfies*

$$\frac{1}{n}\|X(\widehat{\beta} - \beta^*)\|_2^2 = O_p\left(\frac{s^2}{n}(\log p)^{3+\frac{1}{r}}\right)$$

$$\|\widehat{\beta} - \beta^*\|_1 = O_p\left(\frac{s^2}{n}(\log p)^{2+\frac{1}{2r}}\right)$$

$$\|\widehat{\beta} - \beta^*\|_2 = O_p\left(\frac{s^{3/2}}{n}(\log p)^{2+\frac{1}{2r}}\right)$$

- Super-consistency

Admissible Range

- L_1 error bound is $s\sqrt{(\log p)/n}$ under iid cross sectional regressions
  - Prices from RE: $s \log p$; from DB $(\log p)^{\frac{1}{2}+\frac{1}{2r}}$
- Faster than Wijler (2022)'s rates
- In reality C_{DB} is unknown. Admissible λ:

$$\frac{(\log p)^{1+\frac{1}{2r}}}{\lambda} + \lambda s \sqrt{\frac{\log p}{n}} \to 0$$

RE and DB under Transformation

- For a unit root process, $\widehat{\sigma}_j / \sqrt{n} = O_p(1)$
- Sample variance

$$\frac{1}{(\log p)^{1/(4r)}} \overset{\mathrm{p}}{\preccurlyeq} \frac{\widehat{\sigma}^2_{\min}}{n} \leq \frac{\widehat{\sigma}^2_{\max}}{n} \overset{\mathrm{p}}{\preccurlyeq} \log p.$$

For the Gram matrix $\tilde{\Sigma} = D^{-1}\widehat{\Sigma}D^{-1}$:

- RE: $\tilde{\kappa} \overset{\mathrm{p}}{\succcurlyeq} \frac{1}{s(\log p)^{3+\frac{1}{4r}}}$
- DB: $\|n^{-1}\sum_{t=1}^{n}\tilde{X}_{t-1}u_t\|_\infty \leq \frac{1}{\sqrt{n}}\tilde{C}_{\mathrm{DB}}(\log p)^{1+\frac{3}{4r}}$

Slightly slower for Slasso than Plasso due to randomness from $\widehat{\sigma}_j$.

RE and DB under Transformation

- For a unit root process, $\widehat{\sigma}_j / \sqrt{n} = O_p(1)$
- Sample variance

$$\frac{1}{(\log p)^{1/(4r)}} \overset{\mathrm{p}}{\preccurlyeq} \frac{\widehat{\sigma}^2_{\min}}{n} \leq \frac{\widehat{\sigma}^2_{\max}}{n} \overset{\mathrm{p}}{\preccurlyeq} \log p.$$

For the Gram matrix $\tilde{\Sigma} = D^{-1}\widehat{\Sigma}D^{-1}$:

- RE: $\tilde{\kappa} \overset{\mathrm{p}}{\succcurlyeq} \frac{1}{s(\log p)^{3+\frac{1}{4r}}}$
- DB: $\|n^{-1}\sum_{t=1}^{n}\tilde{X}_{t-1}u_t\|_\infty \leq \frac{1}{\sqrt{n}}\tilde{C}_{\mathrm{DB}}(\log p)^{1+\frac{3}{4r}}$

Slightly slower for Slasso than Plasso due to randomness from $\widehat{\sigma}_j$.

Slasso for unit roots

Theorem

Specify $\lambda = \frac{\tilde{C}_{\mathrm{DB}}}{\sqrt{n}}(\log p)^{1+\frac{3}{4r}}$*, and under the Assumptions we have*

$$\begin{aligned}
\frac{1}{n}\|\tilde{X}(\tilde{\beta}-\tilde{\beta}^*)\|_2^2 &= O_p\left(\frac{s^2}{n}(\log p)^{5+\frac{3}{2r}}\right)\\
\|\tilde{\beta}-\tilde{\beta}^*\|_1 &= O_p\left(\frac{s^2}{\sqrt{n}}(\log p)^{4+\frac{1}{r}}\right)\\
\|\tilde{\beta}-\tilde{\beta}^*\|_2 &= O_p\left(\frac{s^{3/2}}{\sqrt{n}}(\log p)^{4+\frac{1}{r}}\right).
\end{aligned}$$

- Remind $\tilde{\beta}_j^* = \widehat{\sigma}_j\beta_j^*$. Super-consistency remains for the original parameter β^*.

Mixed roots

- Pure unit root is a toy model.
- Complex patterns in reality.
- Study a mixture of I(1) and I(0) regressors
- Let $(e_t^\top, Z_t^\top, u_t)^\top = \Phi\varepsilon_t$:

$$
\begin{aligned}
Y_t &= \alpha^* + X_{t-1}^\top \beta^* + Z_{t-1}^\top \gamma^* + u_t \\
&= \alpha^* + \begin{pmatrix} X_{t-1} \\ Z_{t-1} \end{pmatrix}^\top \begin{pmatrix} \beta^* \\ \gamma^* \end{pmatrix} + u_t \\
&= \alpha^* + W_{t-1}^\top \theta^* + u_t
\end{aligned}
$$

- OLS for the original data

$$
\widehat{\theta}^{ols} - \theta^* = (W^\top W)^{-1} W^\top u
$$

Mixed roots

- Pure unit root is a toy model.
- Complex patterns in reality.
- Study a mixture of I(1) and I(0) regressors
- Let $(e_t^\top, Z_t^\top, u_t)^\top = \Phi \varepsilon_t$:

$$\begin{aligned} Y_t &= \alpha^* + X_{t-1}^\top \beta^* + Z_{t-1}^\top \gamma^* + u_t \\ &= \alpha^* + \begin{pmatrix} X_{t-1} \\ Z_{t-1} \end{pmatrix}^\top \begin{pmatrix} \beta^* \\ \gamma^* \end{pmatrix} + u_t \\ &= \alpha^* + W_{t-1}^\top \theta^* + u_t \end{aligned}$$

- OLS for the original data

$$\widehat{\theta}^{ols} - \theta^* = (W^\top W)^{-1} W^\top u$$

OLS for Mixed roots

- Lee, Shi and Gao (2022): Under fixed p, asymptotic distribution of OLS is

$$\begin{pmatrix} n(\widehat{\beta}^{ols} - \beta^*) \\ \sqrt{n}(\widehat{\gamma}^{ols} - \gamma^*) \end{pmatrix} = \begin{pmatrix} \frac{X^\top X}{n^2} & \frac{X^\top Z}{n^{3/2}} \\ \frac{Z^\top X}{n^{3/2}} & \frac{Z^\top Z}{n} \end{pmatrix}^{-1} \begin{pmatrix} \frac{X^\top u}{n} \\ \frac{Z^\top u}{\sqrt{n}} \end{pmatrix}$$

$$\Longrightarrow \begin{pmatrix} ran.mat & 0 \\ 0 & const \end{pmatrix}^{-1} \begin{pmatrix} ran.vec \\ normal \end{pmatrix}$$

Plasso for Mixed Roots

- Admissible λ:
  - I(1) part: $\frac{(\log p)^{1+\frac{1}{2r}}}{\lambda} + \lambda s \sqrt{\frac{\log p}{n}} \to 0$, implies

$$\lambda \succeq (\log p)^{1+\frac{1}{2r}} \to \infty$$

  - I(0) part: $\sqrt{\frac{\log p}{n}} / \lambda + s\lambda \to 0$, implies

$$\lambda \preceq s^{-1} \to 0$$

- Under fixed p, variable selection effect and consistent estimation are incompatible in the two types of regressions.
- Effects are observed in numerical works.

Plasso for Mixed Roots

- Admissible λ:
  - I(1) part: $\frac{(\log p)^{1+\frac{1}{2r}}}{\lambda} + \lambda s \sqrt{\frac{\log p}{n}} \to 0$, implies

$$\lambda \succeq (\log p)^{1+\frac{1}{2r}} \to \infty$$

  - I(0) part: $\sqrt{\frac{\log p}{n}} / \lambda + s\lambda \to 0$, implies

$$\lambda \preceq s^{-1} \to 0$$

- Under fixed p, variable selection effect and consistent estimation are incompatible in the two types of regressions.
- Effects are observed in numerical works.

Slasso for Mixed Roots

- If $\lambda = \frac{\tilde{C}_{\mathrm{DB}}^{w}}{\sqrt{n}}(\log p)^{1+\frac{3}{4r}}$, then the same rates for Slasso above apply to $n^{-1}\|\tilde{W}(\tilde{\theta} - \tilde{\theta}^*)\|_2^2$, $\|\tilde{\theta} - \tilde{\theta}^*\|_1$ and $\|\tilde{\theta} - \tilde{\theta}^*\|_2$.

- Summary:

	Plasso	Slasso
Pure I(1)	consistent	consistent
Mix I(1) and I(0)	inconsistent	consistent

Slasso for Mixed Roots

- If $\lambda = \frac{\tilde{C}^w_{\mathrm{DB}}}{\sqrt{n}} (\log p)^{1+\frac{3}{4r}}$, then the same rates for Slasso above apply to $n^{-1}\|\tilde{W}(\tilde{\theta} - \tilde{\theta}^*)\|_2^2$, $\|\tilde{\theta} - \tilde{\theta}^*\|_1$ and $\|\tilde{\theta} - \tilde{\theta}^*\|_2$.

- Summary:

	Plasso	Slasso
Pure I(1)	consistent	consistent
Mix I(1) and I(0)	inconsistent	consistent

Variable Selection

- Karush-Kuhn-Tucker condition:

$$\frac{2}{n}\check{W}_j^\top \check{u} = \lambda \times \mathrm{sign}(\check{\theta}_j) \quad \text{if } \check{\theta}_j \neq 0$$
$$|\frac{2}{n}\check{W}_j^\top \check{u}| < \lambda \quad \text{if } \check{\theta}_j = 0$$

- More likely to select variables with large s.d.
- Observed in empirical application and simulations.

Section 4

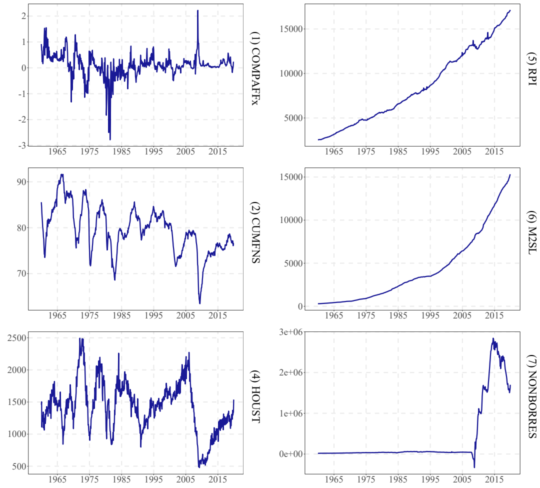
Empirical Application

Our Application: UNRATE

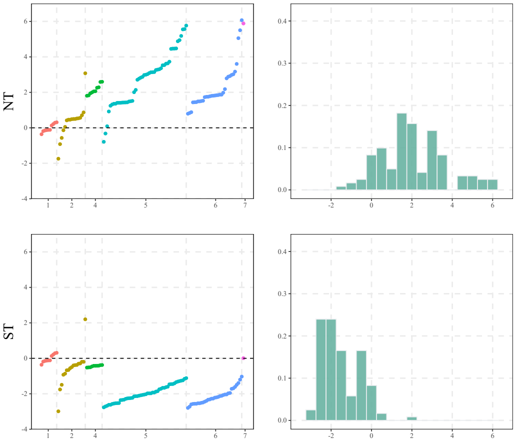
- FRED-MD database
- Data: 1960:Jan–2019:Dec

121 Predictors

- TCODE for stationarity: (1) "nil", (2) Δy_t , (3) $\Delta^2 y_t$, (4) $\log(y_t)$, (5) $\Delta \log(y_t)$, (6) $\Delta^2 \log(y_t)$, (7) $\Delta(y_t/y_{t-1} - 1)$



S.D. of Variables



Note: y-axis is *logarithm base 10*

LASSO Implementation

- Data-driven tuning parameter
  - Block 10-fold cross validation (CV)
- 121 predictors
  - All other variables in database

RMSPE

h	n	Benchmarks		121 Predictors			
				NT		ST	
		RWwD	AR	Plasso	Slasso	Plasso	Slasso
1	120	0.154	0.150	0.639	*0.144*	0.889	0.511
	240	0.154	0.149	0.614	*0.145*	0.632	0.647
	360	0.154	0.144	0.518	*0.150*	1.864	1.920
2	120	0.230	0.214	0.689	*0.195*	0.903	0.536
	240	0.230	0.205	0.821	*0.173*	0.635	0.643
	360	0.229	0.199	0.600	*0.189*	0.744	1.561
3	120	0.306	0.281	0.732	*0.266*	0.953	0.563
	240	0.306	0.262	0.726	*0.242*	0.641	0.654
	360	0.305	0.255	0.654	*0.225*	0.741	1.177

- Slasso better than Plasso
- NT better than ST

Construct More Predictors

- 504 predictors
  - lagged y (Bai and Ng, 2008)
  - 4 factors (Stock and Watson, 2002)
  - $121 + 1 + 4 = 126$
  - $126 \times 4 = 504$

h	n	Benchmarks		121 Predictors				504 Predictors			
				NT		ST		NT		ST	
		RWwD	AR	Plasso	Slasso	Plasso	Slasso	Plasso	Slasso	Plasso	Slasso
1	120	0.154	0.150	0.639	*0.144*	0.889	0.511	0.578	***0.141***	0.470	0.148
	240	0.154	0.149	0.614	*0.145*	0.632	0.647	0.766	***0.129***	0.239	0.134
	360	0.154	0.144	0.518	*0.150*	1.864	1.920	0.736	***0.129***	0.192	0.134
2	120	0.230	0.214	0.689	*0.195*	0.903	0.536	0.642	***0.192***	0.548	0.203
	240	0.230	0.205	0.821	*0.173*	0.635	0.643	0.878	***0.165***	0.306	0.176
	360	0.229	0.199	0.600	*0.189*	0.744	1.561	0.753	***0.169***	0.259	0.176
3	120	0.306	0.281	0.732	*0.266*	0.953	0.563	0.710	0.320	0.644	***0.264***
	240	0.306	0.262	0.726	*0.242*	0.641	0.654	1.011	0.218	0.389	***0.212***
	360	0.305	0.255	0.654	*0.225*	0.741	1.177	0.786	***0.212***	0.330	0.218

Selected Variables

Macroeconomic domain knowledge is important for machine learning applications!

- Choose NT or ST
- Choose sets of regressors (lagged y, factor, lagged w ...)

(a) 121 Predictors

n	NT		ST	
	Plasso	Slasso	Plasso	Slasso
120	4.553	16.206	4.833	26.228
240	12.381	22.764	21.275	62.458
360	12.867	32.808	24.092	66.156

(b) 504 Predictors

n	NT		ST	
	Plasso	Slasso	Plasso	Slasso
120	10.428	13.058	4.858	21.150
240	9.494	8.786	3.847	22.958
360	8.542	8.747	3.875	23.933

Section 5

Simulations

DGP Design: Leading Case

- Innovation $v_t = 0.4v_{t-1} + \epsilon_t$, where $\epsilon_t \sim iid\ \mathcal{N}(0, 0.84\Sigma)$, with $\Sigma_{ij} = 0.8^{|j-j'|}$.
- $n \in \{120, 240, 360\}$, $p = 2n$, $p_x = \{0.5n, 0.8n, 1.2n, 1.5n\}$ and $s_x = s_z = 2\lceil \log n \rceil$
- $\gamma^* = (0.3 \times [s_z]^\top, 0_{p_z - s_z}^\top)^\top$

  DGP1 $\theta^*_{(1)} = (\beta^{*\top}_{(1)}, \gamma^{*\top})^\top$

  DGP2 $\theta^*_{(2)} = (\beta^{*\top}_{(2)}, \gamma^{*\top})^\top$
- CV λ
- Calibrated λ

DGP Design: Leading Case

- Innovation $v_t = 0.4v_{t-1} + \epsilon_t$, where $\epsilon_t \sim iid\ \mathcal{N}(0, 0.84\Sigma)$, with $\Sigma_{ij} = 0.8^{|j-j'|}$.
- $n \in \{120, 240, 360\}$, $p = 2n$, $p_x = \{0.5n, 0.8n, 1.2n, 1.5n\}$ and $s_x = s_z = 2\lceil \log n \rceil$
- $\gamma^* = (0.3 \times [s_z]^\top, 0^\top_{p_z - s_z})^\top$

  DGP1 $\theta^*_{(1)} = (\beta^{*\top}_{(1)}, \gamma^{*\top})^\top$

  DGP2 $\theta^*_{(2)} = (\beta^{*\top}_{(2)}, \gamma^{*\top})^\top$
- CV λ
- Calibrated λ

DGP Design: Leading Case

- Innovation $v_t = 0.4v_{t-1} + \epsilon_t$, where $\epsilon_t \sim iid\ \mathcal{N}(0, 0.84\Sigma)$, with $\Sigma_{ij} = 0.8^{|j-j'|}$.
- $n \in \{120, 240, 360\}$, $p = 2n$, $p_x = \{0.5n, 0.8n, 1.2n, 1.5n\}$ and $s_x = s_z = 2\lceil \log n \rceil$
- $\gamma^* = (0.3 \times [s_z]^\top, 0_{p_z - s_z}^\top)^\top$

  DGP1 $\theta^*_{(1)} = (\beta^{*\top}_{(1)}, \gamma^{*\top})^\top$

  DGP2 $\theta^*_{(2)} = (\beta^{*\top}_{(2)}, \gamma^{*\top})^\top$
- CV λ
- Calibrated λ

Outcomes for DGP1

n	p_x	p_z	RMSPE					RMSE for estimated coefficients				
			Oracle	CV λ		Calibrated λ		Oracle	CV λ		Calibrated λ	
				Plasso	Slasso	Plasso	Slasso		Plasso	Slasso	Plasso	Slasso
DGP1												
120	60	180	1.149	1.678	*1.269*	1.527	***1.256***	0.846	1.232	*0.913*	1.108	***0.899***
	96	144	1.139	1.760	*1.253*	1.524	***1.239***	0.846	1.298	*0.906*	1.122	***0.891***
	144	96	1.136	1.791	*1.257*	1.570	***1.245***	0.847	1.316	*0.897*	1.131	***0.882***
	180	60	1.143	1.852	*1.240*	1.561	***1.232***	0.843	1.345	*0.879*	1.132	***0.864***
240	120	360	1.069	2.218	*1.229*	1.546	***1.167***	0.610	1.425	*0.710*	0.968	***0.685***
	192	288	1.071	2.221	*1.219*	1.538	***1.161***	0.612	1.464	*0.710*	0.978	***0.682***
	288	192	1.071	2.261	*1.221*	1.528	***1.159***	0.610	1.522	*0.705*	0.981	***0.673***
	360	120	1.066	2.340	*1.227*	1.569	***1.163***	0.607	1.545	*0.700*	0.985	***0.662***
360	180	540	1.059	2.397	*1.202*	1.531	***1.141***	0.483	1.448	*0.575*	0.867	***0.554***
	288	432	1.048	2.474	*1.211*	1.547	***1.138***	0.478	1.491	*0.569*	0.871	***0.545***
	432	288	1.051	2.531	*1.200*	1.547	***1.132***	0.478	1.536	*0.569*	0.877	***0.539***
	540	180	1.039	2.591	*1.198*	1.554	***1.125***	0.482	1.549	*0.571*	0.880	***0.537***

Variable Selection in Categories

- Plasso makes more mistakes in both active and inactive X than Slasso
    - X variables are more influential
    - Substantial bias in active Z variables
- Slasso keeps balance in both X and Z

DGP Design: Pure Unit Root

- $p_x = \{0.5n, 0.8n, 1.2n, 1.5n\}$ and $s_x = 2\lceil \log n \rceil$

  DGP3 $\theta^*_{(3)} = \beta^{*\top}_{(1)}$.

  DGP4 $\theta^*_{(4)} = \beta^{*\top}_{(2)}$.

Outcomes for DGP3

n	p_x	RMSPE					RMSE for estimated coefficients				
		Oracle	CV λ		Calibrated λ		Oracle	CV λ		Calibrated λ	
			Plasso	Slasso	Plasso	Slasso		Plasso	Slasso	Plasso	Slasso
DGP3											
120	60	1.108	*1.111*	1.127	***1.084***	1.106	0.383	*0.328*	0.350	***0.284***	0.302
	96	1.088	*1.103*	1.109	***1.073***	1.085	0.384	*0.323*	0.347	***0.283***	0.309
	144	1.075	1.133	*1.110*	***1.070***	1.082	0.384	*0.285*	0.323	***0.282***	0.315
	180	1.067	1.135	*1.118*	***1.079***	1.093	0.382	*0.288*	0.325	***0.283***	0.316
240	120	1.042	*1.058*	1.068	***1.044***	1.058	0.228	*0.211*	0.233	***0.195***	0.216
	192	1.062	*1.079*	1.095	***1.063***	1.081	0.226	*0.212*	0.238	***0.196***	0.221
	288	1.046	1.140	*1.089*	***1.056***	1.072	0.226	*0.206*	0.231	***0.196***	0.226
	360	1.046	1.155	*1.104*	***1.070***	1.084	0.226	*0.207*	0.234	***0.197***	0.229
360	180	1.024	*1.043*	1.051	***1.033***	1.042	0.151	*0.155*	0.176	***0.146***	0.166
	288	1.031	*1.054*	1.079	***1.045***	1.064	0.150	*0.157*	0.181	***0.147***	0.171
	432	1.037	1.142	*1.082*	***1.050***	1.065	0.149	*0.161*	0.178	***0.148***	0.174
	540	1.024	1.122	*1.066*	***1.035***	1.052	0.150	*0.162*	0.181	***0.149***	0.178

Conclusion

- LASSO in high dimensional predictive regression
- RE and DB
- Plasso vs. Slasso
    - Plasso possesses smaller error bounds for pure unit roots
    - Slasso enjoys theoretical guarantees and better numerical performances for mixed roots
- Extensions
    - Inference
    - Local unit roots and cointegrated predictors
    - Other machine learning methods

Conclusion

- LASSO in high dimensional predictive regression
- RE and DB
- Plasso vs. Slasso
  - Plasso possesses smaller error bounds for pure unit roots
  - Slasso enjoys theoretical guarantees and better numerical performances for mixed roots

- Extensions
  - Inference
  - Local unit roots and cointegrated predictors
  - Other machine learning methods