Sparse Models for Dependent Data

Marcelo C. Medeiros

Department of Economics
Pontifical Catholic University of Rio de Janeiro (PUC-Rio)

Lecture 2
Linear Time-Series Regression Model

- We are interested in linear models
  \[
  y_t = \theta_1 x_{1t} + \cdots + \theta_n x_{nt} + u_t, \\
  = \theta' x_t + u_t
  \]
  
  where
  - \( y_t \) is univariate,
  - \( x_t = (x_{1t}, \ldots, x_{nt})' \) may contain lags of \( y_t \),
  - \( x_t \) weakly exogenous,
  - \( u_t \) is a zero-mean innovation,
  - \( m \) irrelevant variables and \( q \) relevant ones.

- General results for vector autoregressive (VAR) will be considered later.

- Applications for instrumental variable (IV) estimation will be also considered.
Penalized Least Squares

A Penalized Least Squares estimator $\hat{\theta}$:

$$
\hat{\theta}(\lambda) = \arg\min_{\theta \in \Theta} \sum_{t=1}^{T} (y_t - \theta'x_t)^2 + \sum_{i=1}^{n} p_{\lambda}(|\theta_i|),
$$

where

- $p_{\lambda}(|\theta_i|)$ is a non-negative penalty function indexed by the regularization parameter $\lambda$. (e.g., $p_{\lambda}(|\theta_i|) = \lambda|\theta_i|^2$, or $p_{\lambda}(|\theta_i|) = \lambda|\theta_i|$).

- $\lambda$ controls the “number of parameters” in the model.

- If $\lambda = \infty$ no variables enter the model, if $\lambda = 0$ it is just the Ordinary Least Squares (OLS) estimator.
A Penalized Least Squares estimator possesses the **oracle property** if

1. selects the “correct” subset of relevant variables; and
2. has the same asymptotic distribution as the OLS estimator as if the “correct” subset of relevant variables were known beforehand.

Such OLS estimator is denoted **oracle estimator**.
Sign Consistency

- Write

\[
\text{sgn}(x) = \begin{cases} 
-1 & , x < 0 \\
0 & , x = 0 \\
1 & , x > 0 
\end{cases}
\]

**Definition (Sign Consistency)**

We say that an estimate \( \hat{\theta} \) is **sign consistent to** \( \theta \) if

\[
P \left( \text{sgn}(\hat{\theta}) = \text{sgn}(\theta) \right) \to 1,
\]

element-wise as \( T \to \infty \).

- It is equivalent to model selection consistency, but NOT estimation consistency.
The Ridge Regression

- The Ridge estimator is defined as follows:

\[
\hat{\theta}_{\text{Ridge}}(\lambda) = \arg\min_{\theta \in \Theta} \sum_{t=1}^{T} (y_t - \theta' x_t)^2 + \lambda \sum_{i=1}^{n} \theta_i^2
\]

- Advantages:
  - “shrinks” towards zero parameters associated with redundant predictors (not exactly);
  - $\lambda$ is a shrinkage parameter to be chosen;
  - the Ridge solution $\hat{\theta}_{\text{Ridge}}$ is easy to find as the problem remains quadratic in $\theta$:

\[
\hat{\theta}_{\text{Ridge}}(\lambda) = (XX' + \lambda I)^{-1} Xy.
\]

- Drawbacks:
  - Not consistent in general; and
  - Provide biased estimators for the non-zero parameters
The LASSO - Tibshirani (JRRS B, 1996)

- Least Absolute Shrinkage and Selection Operator:

\[ \hat{\theta}_{\text{LASSO}}(\lambda) = \arg \min_{\theta \in \Theta} \sum_{t=1}^{T} (y_t - \theta' x_t)^2 + \lambda \sum_{i=1}^{n} |\theta_i| \]

- Advantages:
  - “Shrinks” to zero parameters associated with redundant predictors;
  - The regularization path can be efficiently estimated;
  - Can handle (many) more variables than observations \((n >> T)\); and
  - Under some conditions can select the correct subset of relevant variables.

- Drawbacks:
  - Not consistent in general; and
  - Provide biased estimators for the non-zero parameters.
The Adaptive LASSO - Zou (JASA, 2006)

- The Adaptive LASSO (adaLASSO) estimator is given by

\[ \hat{\theta}_{adaLASSO} = \arg \min_{\theta \in \Theta} \sum_{t=1}^{T} (y_t - \theta' x_t)^2 + \lambda \sum_{i=1}^{n} w_i |\theta_i|, \]

where \( w_1, \ldots, w_n \) are non-negative pre-defined weights.

- Properties:
  - usually \( w_i = |\hat{\theta}_i^{(ols)}|^{-\tau} \), for \( 0 < \tau \leq 1 \);
  - redundant variables have larger weights;
  - consistent under milder condition;
  - provide consistent estimates for the non-zero parameters;
  - has the oracle property.
Consider the following data generating process (DGP):

\[ y_t = 0.7y_{t-1} + 0.5x_{1,t-1} + 0.9x_{2,t-1} + \varepsilon_t, \quad \varepsilon_t \sim \text{NID}(0,1) \]

\[ x_t = f_t + u_t, \quad u_t \sim \text{NID}(0, I) \]

\[ f_t = 0.8f_{t-1} + \nu_t, \quad \nu_t \sim \text{NID}(0,1) \]

Different forecasts:
1. OLS with increasing number of regressors;
2. OLS with the correct number of regressors (ORACLE);
3. DFM with the true factor;
4. DFM with the estimated factor (1st PC);
5. LASSO (penalty selected by the BIC);
6. adaLASSO (penalty selected by the BIC);
Forecasting with Many Predictors

- $T = 100, N = 95$; 100 observations out-of-sample.

![Graph showing forecasting error variance vs. number of included regressors for different models: All variables, Factor, Estimated factor, LASSO, adaLASSO. The graph illustrates the variance decreasing as the number of included regressors increases, with distinct lines for each model showing varying levels of variance reduction.]
Forecasting with Many Predictors

N = 150 variables; T = 100 observations; 100 observations out-of-sample

- Oracle
- Factor
- Estimated factor
- LASSO
- adaLASSO

Marcelo C. Medeiros
Sparse Models for Dependent Data
Forecasting with Many Predictors

N = 550 variables; T = 500 observations; 100 observations out-of-sample

- Oracle
- Factor
- Estimated factor
- LASSO
- adaLASSO

Marcelo C. Medeiros
Sparse Models for Dependent Data
Forecasting with Many Predictors

N = 1000 variables; T = 500 observations; 100 observations out-of-sample
Inference with Many Controls

Set $\beta = 1$ and consider the following DGP:

$$y_t = \beta x_{1,t} + 0.5 x_{2,t} + 0.9 x_{3,t} + \varepsilon_t, \quad \varepsilon_t \sim \text{NID}(0, 1)$$

$$x_t = f_t + u_t, \quad u_t \sim \text{NID}(0, I)$$

$$f_t = 0.8 f_{t-1} + \nu_t, \quad \nu_t \sim \text{NID}(0, 1)$$

Our interest is to estimate $\beta$. However, we do not know the correct set of controls to include in the equation.

Different estimates:

1. OLS with no controls;
2. OLS with the correct controls (ORACLE);
3. OLS with the true factor;
4. OLS with the estimated factor (1st PC);
5. LASSO (penalty selected by the BIC);
6. adaLASSO (penalty selected by the BIC);
Inference with Many Controls

$N = 150$ variables; $T = 100$ observations

- No controls
- Oracle
- Factor
- Estimated factor
- LASSO
- adaLASSO
Inference with Many Controls

\[ N = 150 \text{ variables}; \ T = 200 \text{ observations} \]
Inference with Many Controls

N = 150 variables; T = 500 observations

- No controls
- Oracle
- Factor
- Estimated factor
- LASSO
- adaLASSO
Inference with Many Controls

N = 150 variables; T = 2000 observations

- No controls
- Oracle
- Factor
- Estimated factor
- LASSO
- adaLASSO

Marcelo C. Medeiros
Sparse Models for Dependent Data
Inference with Many Controls

N = 150 variables; T = 500 observations; OLS weights

- No controls
- Oracle
- Factor
- Estimated factor
- LASSO
- adaLASSO

Marcelo C. Medeiros
Sparse Models for Dependent Data
Quarterly data: 1959:1 – 2011.1

Dependent variable \((\pi_t)\): % changes in the GDP deflator (first vintage).

Explanatory variables \((x_t)\): 47 variables and one lag of inflation.

Forecasting equation:

\[
\pi_{t+1} = \theta' x_t + u_{t+1}.
\]

Four set of models: benchmark, forecast combination, penalized regressions, AR + two factors.
Table: Forecasting Results: Out-of-sample $R^2 \times 100$.

<table>
<thead>
<tr>
<th>Initial Year</th>
<th>1970</th>
<th>1983</th>
<th>1997</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Benchmark Models:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Regressors</td>
<td>25.29</td>
<td>19.47</td>
<td>29.29</td>
</tr>
<tr>
<td>AR(1)</td>
<td>79.54</td>
<td>78.09</td>
<td>78.60</td>
</tr>
<tr>
<td><strong>Forecast Combination:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>37.45</td>
<td>1.22</td>
<td>18.68</td>
</tr>
<tr>
<td>Median</td>
<td>17.47</td>
<td>2.65</td>
<td>8.88</td>
</tr>
<tr>
<td>Trimmed Mean</td>
<td>36.09</td>
<td>0.45</td>
<td>17.16</td>
</tr>
<tr>
<td>DMSPE (0.2)</td>
<td>70.68</td>
<td>50.43</td>
<td>54.37</td>
</tr>
<tr>
<td>DMSPE (0.5)</td>
<td>67.75</td>
<td>42.45</td>
<td>46.80</td>
</tr>
<tr>
<td>DMSPE (0.9)</td>
<td>60.97</td>
<td>29.68</td>
<td>38.18</td>
</tr>
<tr>
<td>DMSPE (1.0)</td>
<td>55.43</td>
<td>24.21</td>
<td>38.55</td>
</tr>
</tbody>
</table>
### Inflation Forecasting - Medeiros and Mendes (2012)

**Table:** Forecasting Results: Out-of-sample $R^2$ ($\times 100$).

<table>
<thead>
<tr>
<th>Initial Year</th>
<th>1970</th>
<th>1983</th>
<th>1997</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LASSO, adaLASSO, and Ridge:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ridge (mean)</td>
<td>74.04</td>
<td>88.68</td>
<td>91.49</td>
</tr>
<tr>
<td>LASSO (mean)</td>
<td>80.51</td>
<td>90.11</td>
<td>93.01</td>
</tr>
<tr>
<td>adaLASSO (mean)</td>
<td>80.32</td>
<td>90.11</td>
<td>93.02</td>
</tr>
<tr>
<td>Ridge (median)</td>
<td>78.32</td>
<td>89.52</td>
<td>92.63</td>
</tr>
<tr>
<td>LASSO (median)</td>
<td>82.00</td>
<td>90.23</td>
<td>93.30</td>
</tr>
<tr>
<td>adaLASSO (median)</td>
<td>80.32</td>
<td>90.11</td>
<td>93.02</td>
</tr>
<tr>
<td>Ridge (trimmed mean)</td>
<td>74.38</td>
<td>88.86</td>
<td>91.73</td>
</tr>
<tr>
<td>LASSO (trimmed mean)</td>
<td>80.65</td>
<td>90.27</td>
<td>93.22</td>
</tr>
<tr>
<td>adaLASSO (trimmed mean)</td>
<td>80.47</td>
<td>90.27</td>
<td>93.23</td>
</tr>
</tbody>
</table>
Table: Forecasting Results: Out-of-sample $R^2 (\times 100)$.

<table>
<thead>
<tr>
<th>Initial Year</th>
<th>1970</th>
<th>1983</th>
<th>1997</th>
</tr>
</thead>
<tbody>
<tr>
<td>LASSO, adaLASSO, and Ridge (cont.):</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ridge (BIC)</td>
<td>83.53</td>
<td>83.42</td>
<td>85.02</td>
</tr>
<tr>
<td>LASSO (BIC)</td>
<td>86.92</td>
<td>88.01</td>
<td>90.42</td>
</tr>
<tr>
<td>adaLASSO (BIC)</td>
<td>85.87</td>
<td>88.01</td>
<td>90.42</td>
</tr>
<tr>
<td>Ridge (Bag BIC)</td>
<td>84.23</td>
<td>89.37</td>
<td>91.70</td>
</tr>
<tr>
<td>LASSO (Bag BIC)</td>
<td>86.52</td>
<td>88.70</td>
<td>90.95</td>
</tr>
<tr>
<td>adaLASSO (Bag BIC)</td>
<td>85.52</td>
<td>88.85</td>
<td>90.95</td>
</tr>
<tr>
<td>PCA (2 components):</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AR(1) + PCA</td>
<td>76.62</td>
<td>69.36</td>
<td>73.75</td>
</tr>
</tbody>
</table>
Returns Forecasting - Medeiros and Mendes (2012)

- Same data as in Welch and Goyal (RFS, 2008).
- Predictive regression:

\[ r_t^* = r_t - r_{f,t} = \alpha_0 + \theta' x_{t-1} + u_t, \quad (1) \]

where \( r_t^* \) represents the market returns in excess to the risk-free rate \( (r_{f,t}) \), \( x_{t-1} \) is a set of lagged predictors, and \( u_t \) is the error term.
Stock returns are measured as continuously compounded returns on the S&P 500 index, including dividends, and the Treasury bill rate is used to compute the equity premium.

With respect to the economic variables used to predict the equity premium, we consider the 14 variables + the lagged return:

Dividend-price ratio (log), $D/P$; Dividend yield (log), $D/Y$; Earnings-price ratio (log), $E/P$; Dividend-payout ratio (log), $D/E$; Stock variance, $SVAR$; Book-to-market ratio, $B/M$; Net equity expansion, $NTIS$; Treasury bill rate, $TBL$; Long-term yield, $LTY$; Long-term return, $LTR$; Term spread, $TMS$; Default yield spread, $DFY$; Default return spread, $DFR$; and Inflation, $INFL$. 
Table: Forecasting Results: Out-of-sample $R^2 \times 100$.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AR(1)</td>
<td>-0.16</td>
<td>-0.10</td>
<td></td>
<td>0.85</td>
</tr>
<tr>
<td>All Regressors</td>
<td>-0.07</td>
<td>-0.81</td>
<td></td>
<td>-1.45</td>
</tr>
<tr>
<td>Dividend Price Ratio</td>
<td>0.31</td>
<td>-0.82</td>
<td></td>
<td>3.80</td>
</tr>
<tr>
<td>Dividend Yield</td>
<td>0.40</td>
<td>-0.81</td>
<td></td>
<td>4.14</td>
</tr>
<tr>
<td>Earning Price Ratio</td>
<td>0.53</td>
<td>0.39</td>
<td></td>
<td>4.33</td>
</tr>
<tr>
<td>Dividend Payout Ratio</td>
<td>-0.51</td>
<td>-1.09</td>
<td></td>
<td>-0.53</td>
</tr>
<tr>
<td>Stock Variance</td>
<td>-0.19</td>
<td>0.39</td>
<td></td>
<td>4.76</td>
</tr>
<tr>
<td>Book to Market</td>
<td>-0.82</td>
<td>-0.71</td>
<td></td>
<td>1.30</td>
</tr>
<tr>
<td>Net Equity Expansion</td>
<td>-0.82</td>
<td>-0.66</td>
<td></td>
<td>-3.20</td>
</tr>
<tr>
<td>T-Bill Rate</td>
<td>-0.73</td>
<td>-2.86</td>
<td></td>
<td>-3.03</td>
</tr>
<tr>
<td>Long Term Yield</td>
<td>-0.76</td>
<td>-2.05</td>
<td></td>
<td>-0.87</td>
</tr>
<tr>
<td>Long Term Spread</td>
<td>0.23</td>
<td>-0.81</td>
<td></td>
<td>-0.93</td>
</tr>
<tr>
<td>Term Spread</td>
<td>-0.96</td>
<td>-2.59</td>
<td></td>
<td>-3.38</td>
</tr>
<tr>
<td>Default Yield Spread</td>
<td>-0.66</td>
<td>-0.66</td>
<td></td>
<td>-0.96</td>
</tr>
<tr>
<td>Default Return Spread</td>
<td>-0.20</td>
<td>-0.03</td>
<td></td>
<td>1.02</td>
</tr>
<tr>
<td>Inflation</td>
<td>0.72</td>
<td>-0.09</td>
<td></td>
<td>-3.97</td>
</tr>
</tbody>
</table>
**Table:** Forecasting Results: Out-of-sample $R^2$ (\times 100).

<table>
<thead>
<tr>
<th>Initial Year</th>
<th>1965</th>
<th>1976</th>
<th>2000</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Forecast Combination:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>1.31</td>
<td>0.54</td>
<td>1.07</td>
</tr>
<tr>
<td>Median</td>
<td>0.92</td>
<td>0.11</td>
<td>0.24</td>
</tr>
<tr>
<td>Trimmed Mean</td>
<td>1.31</td>
<td>0.54</td>
<td>1.07</td>
</tr>
<tr>
<td>DMSPE (0.2)</td>
<td>1.37</td>
<td>0.47</td>
<td>0.94</td>
</tr>
<tr>
<td>DMSPE (0.5)</td>
<td>1.36</td>
<td>0.45</td>
<td>1.14</td>
</tr>
<tr>
<td>DMSPE (0.9)</td>
<td>1.32</td>
<td>0.49</td>
<td>1.18</td>
</tr>
<tr>
<td>DMSPE (1.0)</td>
<td>1.30</td>
<td>0.50</td>
<td>1.12</td>
</tr>
</tbody>
</table>
**Table:** Forecasting Results: Out-of-sample $R^2$ ($\times 100$).

<table>
<thead>
<tr>
<th>Initial Year</th>
<th>1965</th>
<th>1976</th>
<th>2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>LASSO, adaLASSO, and Ridge:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ridge (mean)</td>
<td>7.91</td>
<td>6.82</td>
<td>12.33</td>
</tr>
<tr>
<td>LASSO (mean)</td>
<td>8.50</td>
<td>7.58</td>
<td>13.36</td>
</tr>
<tr>
<td>adaLASSO (mean)</td>
<td>8.50</td>
<td>7.58</td>
<td>13.36</td>
</tr>
<tr>
<td>Ridge (median)</td>
<td>8.65</td>
<td>7.36</td>
<td>13.91</td>
</tr>
<tr>
<td>LASSO (median)</td>
<td>9.20</td>
<td>8.20</td>
<td>14.17</td>
</tr>
<tr>
<td>adaLASSO (median)</td>
<td>8.50</td>
<td>7.58</td>
<td>13.36</td>
</tr>
<tr>
<td>Ridge (trimmed mean)</td>
<td>7.96</td>
<td>6.84</td>
<td>12.42</td>
</tr>
<tr>
<td>LASSO (trimmed mean)</td>
<td>8.57</td>
<td>7.64</td>
<td>13.49</td>
</tr>
<tr>
<td>adaLASSO (trimmed mean)</td>
<td>8.57</td>
<td>7.64</td>
<td>13.49</td>
</tr>
</tbody>
</table>
Returns Forecasting - Medeiros and Mendes (2012)

Table: Forecasting Results: Out-of-sample $R^2$ ($\times 100$).

<table>
<thead>
<tr>
<th>Initial Year</th>
<th>1965</th>
<th>1976</th>
<th>2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>LASSO, adaLASSO, and Ridge (cont.):</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ridge (BIC)</td>
<td>7.36</td>
<td>6.95</td>
<td>13.13</td>
</tr>
<tr>
<td>LASSO (BIC)</td>
<td>7.36</td>
<td>6.95</td>
<td>13.11</td>
</tr>
<tr>
<td>adaLASSO (BIC)</td>
<td>7.36</td>
<td>6.95</td>
<td>13.11</td>
</tr>
<tr>
<td>Ridge (Bag BIC)</td>
<td>6.02</td>
<td>5.72</td>
<td>10.73</td>
</tr>
<tr>
<td>LASSO (Bag BIC)</td>
<td>6.07</td>
<td>5.76</td>
<td>10.80</td>
</tr>
<tr>
<td>adaLASSO (Bag BIC)</td>
<td>6.07</td>
<td>5.76</td>
<td>10.80</td>
</tr>
<tr>
<td>PCA (2 components):</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AR(1) + PCA</td>
<td>1.30</td>
<td>-0.34</td>
<td>5.53</td>
</tr>
</tbody>
</table>
How much $R^2$? (CT 2008)

- Excess return: $y_{t+1} = \mu + \beta x_t + \varepsilon_{t+1} := \mu + z_t + \varepsilon_{t+1}$
- Investor with single-period horizon, mean-variance preferences

$$U = \text{expected return} - \frac{\gamma}{2} \text{ portfolio variance.}$$

- **Without observing** $x_t$: Portfolio weight

$$w = \frac{1}{\gamma \sigma_z^2 + \sigma_\varepsilon^2} \cdot \frac{\mu}{\sigma_z^2 + \sigma_\varepsilon^2}.$$

Average EP

$$AEP_0 := \frac{1}{\gamma} \frac{\mu^2}{\sigma_z^2 + \sigma_\varepsilon^2} = \frac{1}{\gamma} S^2,$$

$S$ Sharpe-ratio.
How much $R^2$? (CT 2008)

- **With observing** $x_t$: Portfolio weight

$$w_t = \frac{1}{\gamma} \frac{\mu + z_t}{\sigma^2_\epsilon}.$$ 

Average EP

$$AEP_1 := \frac{1}{\gamma} \frac{\mu^2 + \sigma^2_z}{\sigma^2_\epsilon} = \frac{1}{\gamma} \frac{S^2 + R^2}{1 - R^2}, \text{ where } R^2 = \frac{\sigma^2_z}{\sigma^2_z + \sigma^2_\epsilon}.$$ 

- Increase in AEP from observing $x_t$:

$$AEP_1 - AEP_0 = \frac{1}{\gamma} \left( \frac{1 + S^2}{1 - R^2} \right) R^2 \approx \frac{1}{\gamma} R^2$$

$$\frac{AEP_1}{AEP_0} = \frac{S^2 + R^2}{(1 - R^2)S^2} \approx \frac{S^2 + R^2}{S^2} = 1 + \frac{R^2}{S^2}$$
How much $R^2$? (CT 2008)

$$\frac{AEP_1}{AEP_0} \approx 1 + \frac{R^2}{S^2}$$

- If $R^2$ is large w.r.t. $S^2$, then an investor can use the information in the predictive regression to obtain a large proportional increase in return.
- CT data set (monthly 1871–2005): $S^2 = 0.0120$.
- Out-of-sample $R^2$ for $d/p$ is 0.0074.
  \[ \frac{R^2}{S^2} = \frac{0.0074}{0.0120} = 0.62 \text{ or } 62\% \]
- Out-of-sample $R^2$ for $e/p$ is 0.0038.
  \[ \frac{R^2}{S^2} = \frac{0.0038}{0.0120} = 0.32 \text{ or } 32\% \]
Let’s consider the LASSO estimator.

\( n = T \) and

\[
\frac{1}{T}X'X = I_{n \times n}.
\]

The LASSO estimator is the soft threshold estimator:

\[
\hat{\theta}_i(\lambda) = \text{sign}(Z_j) \left( |Z_j| - \frac{\lambda}{2} \right)_+,
\]

where \((x)_+ = \max(x, 0)\) and \(Z_j = \frac{X'Y}{T}\) equals the OLS estimator for \(\theta_j\).
Asymptotic Theory

- Let’s write the model in matrix form:

\[ \mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{U}. \]

- The number of variables may depend on the sample size 
  \( n \equiv n_T \) (triangular array) and \( n \gg T \).

- Linear model with zero-mean IID errors and orthogonal to the 
  set of regressors.

- A consistency result requires the following sparsity assumption

\[ \|\boldsymbol{\theta}_0\|_1 = \|\boldsymbol{\theta}_0, T\|_1 = o\left(\sqrt{\frac{T}{\log(n)}}\right). \]
Under some regularity conditions on the error term, assuming

\[ \|\theta_0\| = o \left( \frac{T}{\log(T)} \right)^{1/4} \]

and for a range of \( \lambda \approx \sqrt{\log(n)/T} \), the LASSO is consistent for estimating the regression function:

\[
\begin{align*}
\left[ \hat{\theta}(\lambda) - \theta_0 \right] \Omega \left[ \hat{\theta}(\lambda) - \theta_0 \right]' &= o_p(1), \text{ as } T \to \infty,
\end{align*}
\]

where \( \Omega = \frac{1}{T} X'X \) in the case of fixed design or \( \Omega \) is the covariance matrix of \( X \) in the random design.
Under certain compatibility condition (or restricted eigenvalue) on $\mathbf{X}$ and for a $\lambda$ in a suitable range of the order $\sqrt{\log(n)/T}$, it is possible to show the oracle inequality for fixed design:

$$
\mathbb{E} \left[ \left\| \mathbf{X} \left( \hat{\theta}(\lambda) - \theta_0 \right) \right\|_2^2 / T \right] = O \left( \frac{s_0 \log(n)}{n \phi^2} \right),
$$

where $s_0 = \text{card}(S_0)$ and $S_0 = \{ j : \theta_{j,0} \neq 0, j = 1, \ldots, n \}$ (active set); $\phi^2$ is the compatibility constant.

A similar result holds for random design as well.
Under compatibility assumptions on $\mathbf{X}$ and on the sparsity $s_0$, it can be shown that for a $\lambda$ in a suitable range of order $\lambda \approx \sqrt{\log(n)/T}$ that

$$\|\hat{\theta}(\lambda) - \theta_0\|_\ell \overset{p}{\to} 0,$$

where $\ell = 1, 2$.

Screening estimator: $\hat{S}(\lambda) = \{j : \hat{\theta}_j(\lambda) \neq 0, j = 1, \ldots, n\}$.

Lest ambitious goal: find at least the covariates which have large coefficients in absolute values

$$S_{0}^{\text{Relevant}}(C) = \{j : |\theta_j| \geq C, j = 1, \ldots, n\}.$$
Asymptotic Theory

Assume that

\[ \| \hat{\theta}_T(\lambda_T) - \theta_0 \|_1 \leq a_T \] with high probability.

Then, under some regularity conditions and for \( C_T > a_T \), with high probability \( \hat{S}_T(\lambda_T) \supset S_0^{\text{Relevant}(C)} \) (variable screening).
For any $t$, $x_t = [x_t(1)', x_t(2)']'$ and $X = [X(1), X(2)]$, where $X(1)$ is the $(T \times q)$ partition with the relevant variables and $X(2)$ is the $(T \times m)$ partition with the irrelevant ones. Write $\theta = [\theta(1)', \theta(2)']'$ where $\theta(1) \in \mathbb{R}^q$ and $\theta(2) \in \mathbb{R}^m$. $\theta_0$ is the true parameter, where $\theta_0 = [\theta_0(1)', 0']'$, with $\theta_0(1) \neq 0$.

Write $\hat{\Omega} = \frac{x'x}{T}$, $\hat{\Omega}_{11} = \frac{x(1)'x(1)}{T}$, $\hat{\Omega}_{22} = \frac{x(2)'x(2)}{T}$ and $\hat{\Omega}_{21} = \hat{\Omega}_{12} = \frac{x(2)'x(1)}{T}$. Set $sgn_0 = sgn(\theta_0(1))$, where $sgn_0j = 1$ if $\theta_0j > 0$ and $sgn_0j = -1$ otherwise.
The irrepresentable condition (IC) is defined as

\[ \left\| \Omega_{21} \Omega_{11}^{-1} \text{sgn}_0 \right\|_\infty \leq \delta, \]

for some \( 0 < \delta < 1 \).

Under IC and the condition

\[ \inf_{j \in S_0} |\theta_{0,j}| >> \sqrt{\frac{s_0 \log(n)}{T}}, \]

and for a suitable choice of \( \lambda = \lambda_T >> \frac{\sqrt{nT}}{T} \), then

\[ P[\hat{S}(\lambda) = S_0] \rightarrow 1, \text{ as } T \rightarrow \infty. \]
Recalling the adaLASSO estimator:

$$
\hat{\theta}_{\text{adaLASSO}}(\lambda) = \arg \min_{\theta} \left( \|Y - X\theta\|^2 / T + \lambda \sum_{i=1}^{n} \left| \frac{\theta_i}{\theta^*_i} \right|^\gamma \right)
$$

It is clear that $\theta^*_i = 0 \Rightarrow \hat{\theta}_{\text{adaLASSO},j} = 0$.

- OLS as initial estimator whenever possible.
- Variable selection follows under the weighted irrepresentable condition:

$$
\left\| \Omega_{21} \Omega_{11}^{-1} W(1) \text{sgn}_0 \right\|_\infty \leq \delta,
$$

where $W(1)$ is diagonal matrix with the weights of the relevant variables.
What We Do?

1. consider “very broad” class of data generating processes;
2. allow for high-dimensional problems \( n = O(T^\alpha) \), for some \( \alpha > 0 \);
3. weaken previous conditions in the literature;
4. show model selection consistency and oracle property under these weak conditions; and
5. show the method works well in simulations, specially for prediction.
Four Sets of Assumptions

2. **REG**: Conditions on the rate of increase of: the total number redundant variables, the total number of relevant parameters, and the size of the regularization parameter.
3. **PAR**: Assumptions regarding the true parameter vector and parameter space.
4. **WIC**: Regularity conditions about the weights $w_j$. 
The stochastic process \( \{y_t\} \) is generated by

\[
y_t = \theta_0 x_{1t} + \cdots + \theta_0 x_{nt} + u_t,
\]

where \( \theta_0 = (\theta_{01}, \ldots, \theta_{0q}, 0, \ldots, 0)' \), i.e., the first \( q \) elements are non-zero and the remaining \( m = n - q \) are zero.

Write \( z_t = (y_t, x_t', u_t)' \),

**DGP 1.** \( \{z_t\} \) is a zero-mean weak-stationary process.

**DGP 2.** \( \mathbb{E}[u_t|x_t] = 0. \)

**DGP 3.** For some \( d \geq 1 \) and some positive constant \( c_d \),

\[
\max_{i=1,\ldots,n} \mathbb{E} \left| T^{-1/2} \sum_{t=1}^{T} x_{it} u_t \right|^{2d} \leq c_d.
\]
Assumption DGP: Data Generating Process

- How strong is DGP 3?

Example (Marcinkiewicz-Zygmund inequality)

If $\mathbb{E}[u_t^{2d}] < \infty$ and $\mathbb{E}[x_{it}^{2d}] < \infty$, then, under some tail conditions,

$$\mathbb{E} \left| T^{-1/2} \sum_{t=1}^{T} x_{it} u_t \right|^{2d} \propto \mathbb{E} \left( T^{-1} \sum_{t=1}^{T} (x_{it} u_t)^2 \right)^d.$$

For $d = 1$ and assuming $\mathbb{E}[u_t^2|x_t] = \mathbb{E}[u_t^2]$, we just need finite variances.
Assumption REG: Rates of Increase of $\lambda$, $m$ and $q$

- Let $\lambda \to \infty$ such that
  
  $R0$. (Regularization Parameter) $\lambda = O\left(\sqrt{\log \frac{T}{T}}\right)$;

  $R1$. (Redundant Variables) $m = O\left(\frac{T^{d\xi}}{(\log T)^{d/2}}\right)$, for some fixed $0 < \xi < \tau$;

  $R4$. (Relevant Variables) $q = o\left(T^{\frac{d}{5d+1}}\right)$.

- For “real” High-Dimension, we need $d \times \xi > 1$. 
PAR 1. The true parameter vector $\theta_0$ is an element of an open subset $\Theta \in \mathbb{R}^n$, around the origin.

PAR 2. There exists a constant $\theta_\star$ satisfying $\min_{i=1,...,q} |\theta_{0i}| \geq \theta_\star / q$.

Condition PAR 2 is more relaxed than the classical “beta-min” condition, i.e., the minimum parameter value is greater than a constant, independent of $T$. 
Assumption WIC: Weighted Irrepresentable Condition

- $S_1 =$ Lower bound on the smallest eigenvalue of $T^{-1} \mathbf{X}(1)' \mathbf{X}(1)$ that may decrease with $q$.
- $S_2 =$ Upper bound on the value of the weights associated with the relevant variables, i.e., $w_1, \ldots, w_q$.
- $S_3 =$ Upper bound on the correlation between the redundant and relevant variables.
- $S_4 =$ Lower bound on the value of the weights associated with the redundant variables, i.e., $w_{q+1}, \ldots, w_n$. 
The adaptive LASSO estimator is given by

$$\hat{\theta} = \arg \min_{\theta \in \Theta} \sum_{t=1}^{T} (y_t - \theta' x_t)^2 + \lambda \sum_{i=1}^{n} w_i |\theta_i|,$$

where $w_1, \ldots, w_n$ are given by $w_i = |\hat{\theta}_j^*|^{-\tau}$.

In a (rough) simplified way, we need

$$\min_{i=1,\ldots,q} |\hat{\theta}_i^*| > c_w / q,$$

and

$$\max_{j=q+1,\ldots,n} |\hat{\theta}_j^*| < c / q^{1+a},$$

for $a = (1 + \gamma) / \tau > 1$. 

Assumption WIC: Weighted Irrepresentable Condition
Theorem (Model Selection Consistency)

Under assumptions DGP, PARAM, WIC and REG

\[ P \left( \text{sgn}(\hat{\theta}) = \text{sgn}(\theta_0) \right) \to 1, \]

as \( T \to \infty \).
The Oracle Property

Let $\hat{\theta}_{ols}(1)$ denote the OLS estimator of the regression

$$y_t = \theta_1 x_{1t} + \cdots + \theta_q x_{qt} + u_t,$$

i.e., the Oracle-OLS estimator; and $\hat{\theta}(1)$ the adaLASSO estimator, corresponding to the correct set of non-zero parameters.

**Theorem (Oracle Property)**

Under assumptions DGP, PARAM, WIC and REG, and for some $q$-dimensional vector $\alpha$ with Euclidean norm 1, we have

$$\sqrt{T} \alpha'(\hat{\theta}(1) - \theta_0(1)) = \sqrt{T} \alpha'(\hat{\theta}_{ols}(1) - \theta_0(1)) + o_p(1).$$
Consider the following data generating process (DGP):

\[ y_t = 0.7y_{t-1} + 0.5x_{1,t-1} + 0.9x_{2,t-1} + \frac{h_t^{1/2}}{u_t} \varepsilon_t, \quad \varepsilon_t \sim \text{NID}(0, 1) \]

\[ h_t = 0.0005 + 0.9h_{t-1} + 0.05u_{t-1}^2 \]

\[ x_t = f_t + u_t, \quad u_t \sim \text{NID}(0, I) \]

\[ f_t = 0.8f_{t-1} + v_t, \quad v_t \sim \text{NID}(0, 1) \]
Simulations - FAR + GARCH

![Graph showing simulations of FAR + GARCH model with time on the x-axis and dependent variable on the y-axis.](image)
Simulations - FAR + GARCH

![Graph showing conditional volatility over time]

- **Time:** 0 to 200
- **Conditional Volatility:** 0.006 to 0.02

**References:**
- Medeiros and Mendes (2012)
- Medeiros and Passos (2012)
- Kock (2012)
- Kock and Callot (2012)
Simulations - FAR + GARCH

N=95; T=100

Oracle
LASSO
adaLASSO

N=95; T=100

Oracle
LASSO
adaLASSO

Oracle
LASSO
adaLASSO

Marcelo C. Medeiros
Sparse Models for Dependent Data
Simulations - FAR + GARCH

N=150; T=100

Oracle
LASSO
adaLASSO

Oracle
LASSO
adaLASSO

Oracle
LASSO
adaLASSO

Oracle
LASSO
adaLASSO
Simulations - FAR + GARCH

N=300; T=100

Oracle
LASSO
adaLASSO

Marcelo C. Medeiros
Sparse Models for Dependent Data
Consider the autoregressive time-series of length 1000 generated by

\[ y_t = 0.2y_{t-1} + 0.1y_{t-3} + 0.2y_{t-5} + 0.3y_{t-10} + 0.1y_{t-15} + u_t, \]

where \( u_t \sim \text{NID}(0, .01) \) (i.i.d. with a normal distribution).

The candidate variables are lags of the dependent variable.

Three values for \( n \): 50, 200 and 500.

The effective number of observations \( T \) is 950, 800 and 500.
Simulations - AR

Table: Frequency Each Relevant Lag is Selected

<table>
<thead>
<tr>
<th>n</th>
<th>$\theta_1 = .2$</th>
<th>$\theta_3 = .1$</th>
<th>$\theta_5 = .2$</th>
<th>$\theta_{10} = .3$</th>
<th>$\theta_{15} = .1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>1000</td>
<td>798</td>
<td>1000</td>
<td>1000</td>
<td>767</td>
</tr>
<tr>
<td>200</td>
<td>1000</td>
<td>544</td>
<td>999</td>
<td>1000</td>
<td>547</td>
</tr>
<tr>
<td>500</td>
<td>931</td>
<td>388</td>
<td>971</td>
<td>1000</td>
<td>471</td>
</tr>
</tbody>
</table>
## Simulations - AR

**Table:** Average Number of Correctly Classified Predictors

<table>
<thead>
<tr>
<th>$n$</th>
<th># Corr. NZ (5)</th>
<th># Corr. Z ($n - 5$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>4.565</td>
<td>44.458</td>
</tr>
<tr>
<td>200</td>
<td>4.090</td>
<td>194.431</td>
</tr>
<tr>
<td>500</td>
<td>3.661</td>
<td>492.539</td>
</tr>
</tbody>
</table>
### Simulations - AR

#### Table: Estimated Parameters and Respective Standard Errors

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\theta_1 = 0.2$</th>
<th>$\theta_3 = 0.1$</th>
<th>$\theta_5 = 0.2$</th>
<th>$\theta_{10} = 0.3$</th>
<th>$\theta_{15} = 0.1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.193 (0.032)</td>
<td>0.074 (0.044)</td>
<td>0.199 (0.037)</td>
<td>0.300 (0.032)</td>
<td>0.074 (0.049)</td>
</tr>
<tr>
<td>500</td>
<td>0.168 (0.061)</td>
<td>0.032 (0.053)</td>
<td>0.184 (0.057)</td>
<td>0.285 (0.050)</td>
<td>0.048 (0.059)</td>
</tr>
</tbody>
</table>

#### Table: Oracle Parameters and Respective Standard Errors

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\theta_1 = 0.2$</th>
<th>$\theta_3 = 0.1$</th>
<th>$\theta_5 = 0.2$</th>
<th>$\theta_{10} = 0.3$</th>
<th>$\theta_{15} = 0.1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.198 (0.029)</td>
<td>0.098 (0.028)</td>
<td>0.197 (0.033)</td>
<td>0.296 (0.030)</td>
<td>0.099 (0.031)</td>
</tr>
<tr>
<td>500</td>
<td>0.199 (0.039)</td>
<td>0.098 (0.037)</td>
<td>0.195 (0.042)</td>
<td>0.294 (0.042)</td>
<td>0.097 (0.042)</td>
</tr>
</tbody>
</table>
Table: Estimated Parameters and Respective Standard Errors

<table>
<thead>
<tr>
<th>n</th>
<th>$\theta_1 = 0.2$</th>
<th>$\theta_3 = 0.1$</th>
<th>$\theta_5 = 0.2$</th>
<th>$\theta_{10} = 0.3$</th>
<th>$\theta_{15} = 0.1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.193 (0.032)</td>
<td>0.074 (0.044)</td>
<td>0.199 (0.037)</td>
<td>0.300 (0.032)</td>
<td>0.074 (0.049)</td>
</tr>
<tr>
<td>500</td>
<td>0.168 (0.061)</td>
<td>0.032 (0.053)</td>
<td>0.184 (0.057)</td>
<td>0.285 (0.050)</td>
<td>0.048 (0.059)</td>
</tr>
</tbody>
</table>

Table: Oracle Parameters and Respective Standard Errors

<table>
<thead>
<tr>
<th>n</th>
<th>$\theta_1 = 0.2$</th>
<th>$\theta_3 = 0.1$</th>
<th>$\theta_5 = 0.2$</th>
<th>$\theta_{10} = 0.3$</th>
<th>$\theta_{15} = 0.1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.198 (0.029)</td>
<td>0.099 (0.028)</td>
<td>0.197 (0.033)</td>
<td>0.296 (0.030)</td>
<td>0.099 (0.031)</td>
</tr>
<tr>
<td>500</td>
<td>0.199 (0.039)</td>
<td>0.098 (0.037)</td>
<td>0.195 (0.042)</td>
<td>0.294 (0.042)</td>
<td>0.097 (0.042)</td>
</tr>
</tbody>
</table>
Simulations - AR

### Table: Estimated Parameters and Respective Standard Errors

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\theta_1 = 0.2$</th>
<th>$\theta_3 = 0.1$</th>
<th>$\theta_5 = 0.2$</th>
<th>$\theta_{10} = 0.3$</th>
<th>$\theta_{15} = 0.1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.193 (0.032)</td>
<td>0.074 (0.044)</td>
<td>0.199 (0.037)</td>
<td>0.300 (0.032)</td>
<td>0.074 (0.049)</td>
</tr>
<tr>
<td>500</td>
<td>0.168 (0.061)</td>
<td>0.032 (0.053)</td>
<td>0.184 (0.057)</td>
<td>0.285 (0.050)</td>
<td>0.048 (0.059)</td>
</tr>
</tbody>
</table>

### Table: Oracle Parameters and Respective Standard Errors

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\theta_1 = 0.2$</th>
<th>$\theta_3 = 0.1$</th>
<th>$\theta_5 = 0.2$</th>
<th>$\theta_{10} = 0.3$</th>
<th>$\theta_{15} = 0.1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.198 (0.029)</td>
<td>0.098 (0.028)</td>
<td>0.197 (0.033)</td>
<td>0.296 (0.030)</td>
<td>0.099 (0.031)</td>
</tr>
<tr>
<td>500</td>
<td>0.199 (0.039)</td>
<td>0.098 (0.037)</td>
<td>0.195 (0.042)</td>
<td>0.294 (0.042)</td>
<td>0.097 (0.042)</td>
</tr>
</tbody>
</table>
Table: MSFE of the Adaptive LASSO and Oracle-OLS estimator

<table>
<thead>
<tr>
<th>$n$</th>
<th>AdaLASSO</th>
<th>OLS-Oracle</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.0147</td>
<td>0.0147</td>
</tr>
<tr>
<td>200</td>
<td>0.0138</td>
<td>0.0136</td>
</tr>
<tr>
<td>500</td>
<td>0.0141</td>
<td>0.0139</td>
</tr>
</tbody>
</table>
Consider the Autoregressive Dynamic Lags model given by

\[ y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \sum_{i=1}^{5} (\beta_{i,1} x_{i,t-1} + \beta_{i,2} x_{i,t-2}) + u_t \]

where \( u_t \sim \text{NID}(0, 2) \), with parameters given by

\[
\begin{align*}
\alpha_1 &= 0.4 & \beta_{1,1} &= 4 & \beta_{2,1} &= 3 & \beta_{3,1} &= 2 & \beta_{4,1} &= 1 & \beta_{5,1} &= 0.5 \\
\alpha_2 &= 0.1 & \beta_{1,2} &= 2 & \beta_{2,2} &= 1.5 & \beta_{3,2} &= 1 & \beta_{4,2} &= 0.5 & \beta_{5,2} &= 0.25
\end{align*}
\]

The candidate predictors are 5 lags of the dependent variable and 20 explanatory variables with 5 lags each.
Simulations - ARDL

- The first two relevant predictors are generated from the following VAR(1) model:

\[
\begin{pmatrix}
    x_{1,t} \\
    x_{2,t}
\end{pmatrix}
= \begin{pmatrix}
    0.5 & 0.3 \\
    0.3 & 0.5
\end{pmatrix}
\begin{pmatrix}
    x_{1,t-1} \\
    x_{2,t-1}
\end{pmatrix} + \begin{pmatrix}
    e_1 \\
    e_2
\end{pmatrix}.
\]

- Four of the redundant predictors are also generated from an VAR(1) model:

\[
\begin{pmatrix}
    x_{6,t} \\
    x_{7,t} \\
    x_{8,t} \\
    x_{9,t}
\end{pmatrix}
= \begin{pmatrix}
    0.5 & 0.3 & 0.1 & 0 \\
    0.3 & 0.5 & 0 & 0.1 \\
    0.1 & 0 & 0.5 & 0.3 \\
    0 & 0.1 & 0.3 & 0.5
\end{pmatrix}
\begin{pmatrix}
    x_{6,t-1} \\
    x_{7,t-1} \\
    x_{8,t-1} \\
    x_{9,t-1}
\end{pmatrix} + \begin{pmatrix}
    e_6 \\
    e_7 \\
    e_8 \\
    e_9
\end{pmatrix}.
\]
The remaining predictors are generated from an AR(1) model:

\[ x_{j,t} = a_j x_{j,t-1} + e_j, \quad \text{for } j = 3, 4, 5, 10, \ldots, 20, \]

with autoregressive parameters \( a_j \) randomly drawn from an uniform distribution \( U[0, 0.8] \).

The error terms \( e_1, \ldots, e_{20} \) are mutually independent, each NID(0, 1).
### Simulations - ARDL

**Table:** Number of selected lags of \( y \)

<table>
<thead>
<tr>
<th>( T )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>2.55%</td>
<td>53.05%</td>
<td>35.10%</td>
<td>7.90%</td>
<td>1.00%</td>
<td>0.40%</td>
</tr>
<tr>
<td>150</td>
<td>0.15%</td>
<td>56.50%</td>
<td>43.25%</td>
<td>0.10%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>200</td>
<td>0.00%</td>
<td>50.15%</td>
<td>49.75%</td>
<td>0.10%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
</tbody>
</table>
Simulations - ARDL

**Table:** Fraction of time each correct variable is selected

<table>
<thead>
<tr>
<th>$T$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>100.00%</td>
<td>100.00%</td>
<td>98.65%</td>
<td>78.80%</td>
<td>47.35%</td>
</tr>
<tr>
<td>150</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>96.60%</td>
</tr>
<tr>
<td>200</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>99.60%</td>
</tr>
</tbody>
</table>
### Simulations - ARDL

#### Table: Number of selected lags

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.05%</td>
<td>29.50%</td>
<td>25.50%</td>
<td>23.50%</td>
<td>30.05%</td>
</tr>
<tr>
<td>150</td>
<td>0.00%</td>
<td>76.05%</td>
<td>21.05%</td>
<td>2.50%</td>
<td>0.40%</td>
</tr>
<tr>
<td>200</td>
<td>0.00%</td>
<td>82.55%</td>
<td>15.75%</td>
<td>1.45%</td>
<td>0.25%</td>
</tr>
</tbody>
</table>
Simulations - ARDL

Table: MSFE of the AdaLASSO and Oracle-OLS estimator

<table>
<thead>
<tr>
<th>$T$</th>
<th>AdaLASSO</th>
<th>OLS-Oracle</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>52.514</td>
<td>52.009</td>
</tr>
<tr>
<td>150</td>
<td>56.340</td>
<td>56.203</td>
</tr>
<tr>
<td>200</td>
<td>53.930</td>
<td>54.066</td>
</tr>
</tbody>
</table>
Investor’s Problem

Let \( \mathbf{r}_{t+1}, \mu_t \equiv \mathbb{E}_t[\mathbf{r}_{t+1}] \), and \( \Sigma_t \equiv \mathbb{E}_t[(\mathbf{r}_{t+1} - \mu_t)(\mathbf{r}_{t+1} - \mu_t)'] \) denote, respectively, an \( N \times 1 \) vector of stock returns, the expected returns and covariance matrix conditional based on information up to period \( t \).

The investor’s problem at time \( t \):

\[
\omega^* = \arg \min_{\omega_t} \omega'_t \hat{\Sigma}_t \omega_t \\
\text{s.t.} \quad \omega'_t \hat{\mu} = \mu_{\text{target}} \\
\omega'_t \mathbf{1} = 1 \\
I(\omega_t < 0)' \mathbf{1} \leq 0.3
\]

where \( \omega_t \) is an \( N \times 1 \) vector of portfolio weights, \( \mu_{\text{target}} \) is the target expected rate of returns from \( t \) to \( t + 1 \) and \( \mathbf{1} \) is a \( N \times 1 \) vector of ones.
Investor’s Problem

- The optimal weights $\omega^*$ which solve the optimization problem can be viewed as a function of the target expected return $\mu_{target}$, the conditional expected return $\mu_t$, and the conditional covariance matrix $\Sigma_t$.

- Instead of fixing a value for $\mu_{target}$, we adopt a different route by considering a “moving target”: at each time $t$ we select the value of $\mu_{target}$ which gives the mean-variance portfolio with the highest Sharpe Ratio.
The conditional expected returns are given by the models discussed earlier and can be generally written as

$$\mu_t = A + X_t B,$$

where $X_t$ is a $N \times K$ matrix of $K$ factors, $A$ a $N \times 1$ vector of parameters and $B$ is a $K \times N$ matrix of parameters.

The elements of $X_t$ will differ accordingly to the model specifications discussed earlier.
Transaction Costs

- Let the “hold portfolio” at period $t + 1$ be defined as the portfolio resulted from keeping the stocks from period $t$.
- Let $\omega_{i,t-1}$, $r_{p,t}$, and $r_{i,t}$ be the portfolio weight on stock $i$ at period $t - 1$, the portfolio return from $t - 1$ to $t$ and the return from stock $i$ from $t - 1$ to $t$, respectively.
- The “hold portfolio” weight on stock $i$ at period $t$ is given by:

$$
\omega_{i,t}^h = \omega_{i,t-1} \frac{1 + r_{i,t}}{1 + r_{p,t}}
$$
Let $c_{i,t}$ be the estimated transaction costs of trading stock $i$ at $t$.

The portfolio return net of trading costs is the portfolio return less the absolute change on weight from the “hold portfolio” multiplied by its transaction costs:

$$r_{p,t+1} = \sum_{i=1}^{N} \omega_{i,t} r_{i,t+1} - c_{i,t} |\omega_{i,t} - \omega_{i,t}^h|$$
Several theoretical studies suggest the optimal strategy with transaction costs should consider a no-trade region, given current position.

If the desired portfolio weights (hereafter called “target portfolio”) is inside the no-trade region, it is optimal not to trade.

The intuition of this result lies in the fact there is a first-order loss when trading inside the no-trade region and only a second-order gain.

Motivated by these theoretical results, Brandt, Sana-Clara, and Valkanov (2009) propose a technique which model the no-trade region as an hypersphere and shrinks the target portfolio to the hold portfolio.
Let $\omega^{\text{target}}$ be the optimal portfolio obtained by the Markowitz approach, $\omega^h$ the weights from the “hold portfolio” and $\kappa^2$ the radius of the no-trade region.

If $\sum_{i=1}^{N_t} (\omega_{i,t}^{\text{target}} - \omega_{i,t}^h)^2 / N_t \leq \kappa^2$, the target portfolio is inside the no-trade region at $t$.

Therefore, the optimal policy is keeping the current portfolio.
However, if \( \sum_{i=1}^{N_t} (\omega_{i,t}^{\text{target}} - \omega_{i,t}^{h})^2 / N_t > \kappa^2 \), the optimal policy is to change the weights towards the target portfolio, up to the no-trade region centered in the target weight:

\[
\omega_{i,t} = \alpha_t \omega_{i,t}^{h} + (1 - \alpha_t) \omega_{i,t}^{\text{target}}
\]

\[
\alpha_t = \frac{\kappa \sqrt{N_t}}{\sqrt{\sum_{i=1}^{N_t} (\omega_{i,t}^{\text{target}} - \omega_{i,t}^{h})^2 / N_t}}
\]

This weighted average can be seen as shrinkage of the optimized portfolio to the hold portfolio.
The Dataset

- 183 portfolios from Kenneth French’s website:
  1. Group 1 – 100 portfolios sorted on size (market equity) and book-to-market (ratio of book equity to market equity).
  2. Group 2 – 25 portfolios sorted on size (market equity) and momentum (previous 2-12 returns). The monthly size breakpoints are the NYSE market equity quintiles. The monthly return breakpoints are NYSE quintiles. The portfolios constructed each month include NYSE, AMEX, and NASDAQ stocks.
  4. Group 4 – 10 portfolios sorted on earnings-price (E/P) ratio. Portfolios are formed on E/P at the end of each June using NYSE breakpoints.

- Predictors: Welch and Goyal (2008).
Economic Value

\[
\sum_{a=1}^{A} \frac{\bar{r}_a^p - \Delta_p}{std_a^p} = \sum_{a=1}^{A} \frac{\bar{r}_a^b}{std_a^b}
\]

<table>
<thead>
<tr>
<th>Portfolio</th>
<th>Sharpe-ratio (basis points)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FF 3-factor</td>
<td>-38</td>
</tr>
<tr>
<td>Intercept</td>
<td>154</td>
</tr>
<tr>
<td>Intercept + Lag</td>
<td>196</td>
</tr>
<tr>
<td>LASSO (equal-weighted combination)</td>
<td>386</td>
</tr>
</tbody>
</table>
Oracle Properties of adaLASSO in DF Regressions

Summary of results:

1. adaLASSO is oracle efficient: estimates parameters consistently, selects the correct sparsity pattern, and estimates the coefficients belonging to the relevant variables at the same asymptotic efficiency as if these has been the only variables in the model.

2. Fixed number of regressors $N$ and $N < T$.

Dickey-Fuller (DF) Regression:

$$
\Delta y_t = \rho_0 y_{t-1} + \sum_{i=1}^{p} \beta_{0,i} \Delta y_{t-i} + u_t.
$$
The parameters \( \theta = (\rho, \beta')' \) are estimated by adaLASSO:

\[
\hat{\theta} = \arg \min_{\theta} \Psi_T(\theta) = \sum_{t=1}^{T} \left( \Delta y_t - \rho y_{t-1} - \sum_{i=1}^{p} \beta_{0,i} \Delta y_{t-i} \right)^2 \\
+ \lambda_T \omega_1^\gamma_1 |\rho| + \lambda_T \sum_{i=1}^{p} \omega_2^\gamma_2 |\beta_i|.
\]
Oracle Properties of adaLASSO in DF Regressions

Result 1: Consistent model selection under nonstationarity
Assume that $\rho_0 = 0$ and that $\varepsilon_t \sim \text{IID}(0, \sigma^2)$ with $\mathbb{E}(\varepsilon_t^4) < \infty$.
Then, if $\frac{\lambda T}{T^{1-\gamma_1}} \to \infty$, $\frac{\lambda T}{T^{1/2-\gamma_2/2}} \to \infty$, and $\frac{\lambda T}{T^{1/2}} \to 0$,

a. Consistency: $\left\| S_T \left[ (\hat{\rho}, \hat{\beta}^\prime) - (0, \beta_0^\prime) \right] \right\|_2 = O_p(1)$, where $S_T = \text{diag}(T, \sqrt{T}, \ldots, \sqrt{T})$.

b. Oracle:
   i. $\mathbb{P}(\hat{\rho} = 0) \to 1$ and $\mathbb{P}(\hat{\beta}_{A_c} = 0) \to 1$, where $A$ is the active and its complement is defined as $A_c$.
   ii. $\sqrt{T}(\hat{\beta}_A - \beta_A) \overset{d}{\to} \mathcal{N}(0, \sigma^2 \Sigma_A^{-1})$. 
Oracle Properties of adaLASSO in DF Regressions

**Result 2: Consistent model selection under stationarity**

Assume that $\rho_0 \in (-2, 0)$ and that $\varepsilon_t \sim \text{IID}(0, \sigma^2)$ with $\mathbb{E}(\varepsilon^4_t) < \infty$. Then, if $\frac{\lambda T}{T^{1/2-\gamma/2}} \to \infty$, and $\frac{\lambda T}{T^{1/2}} \to 0$,

a. Consistency: $\left\| \frac{\lambda}{T^{1/2-\gamma/2}} \left[ (\hat{\rho}, \hat{\beta}') - (\rho_0, \beta_0)' \right] \right\|_2 = O_p(1)$.

b. Oracle:

i. $P(\hat{\rho} = 0) \to 0$ and $P(\hat{\beta}_{A_c} = 0) \to 1$.

ii. $\sqrt{T}(\hat{\theta}_A - \theta_A) \overset{d}{\to} N(0, \sigma^2 Q_A^{-1})$. 
Simulations

- Consider the following DGP:

\[ y_t = \delta y_{t-1} + \nu_t \]

\[ \nu_t = 0.6\nu_{t-1} - 0.3\nu_{t-4} + e_t, \quad e_t \sim \text{NID}(0,1). \]

- Two cases: \( \delta = 1 \) and \( \delta = 0 \);
- ADF test with candidate lags.
- \( T = 40, 80, 160 \) observations, 1000 simulations.
Simulations

- LASSO, Ridge, and adaLASSO correctly selects 100% of the cases under the null.
- Under the alternative the “rejection” rates were: 38% (LASSO), 42% (Ridge), and 20% (adaLASSO)
LASSO, Ridge, and adaLASSO correctly select 100% of the cases under the null.

Under the alternative the “rejection” rates were: 40% (LASSO), 45% (Ridge), and 31% (adaLASSO)
The graph shows the empirical size and power for different nominal sizes. The data points indicate the following:

- **LASSO, Ridge, and adaLASSO** correctly selects 100% of the cases under the null.
- Under the alternative, the "rejection" rates were: 71% (LASSO), 74% (Ridge), and 55% (adaLASSO).
Oracle Inequalities for VARs

• Summary of results:
  1. Non-asymptotic oracle inequalities for the prediction error and estimation accuracy of the LASSO in stationary VAR models.
  2. $N \gg T$.

• The model:

$$Y_t = \sum_{i=1}^{p} A_i Y_{t-i} + u_t,$$

where $u_t \sim \text{NID}(0, \Sigma)$.

• Normality is a key condition.
Consider the following VAR:

\[ Z = XC + V, \]
\[ z = (I \otimes X)c + v, \]

where \( Z \) and \( U \) are \((T \times m)\) matrices and \( X \) is a \((T \times k)\) matrix, \( k = m(p + 1) \), \( z \) and \( v \) are \((mT \times 1)\) vectors, \( I_m \) is an identity matrix of dimension \( m \), and \( c = \text{vec}(C) \).

We omit the intercept in \( X \) to simplify the derivations.
Under normality of the errors the likelihood is given as:

\[
\log \mathcal{L}(c, \Sigma_v) \propto |\Sigma_v \otimes I_T|^{-\frac{1}{2}} \times \\
\exp \left\{ -\frac{1}{2} [z - (I_m \otimes Z) c]' \left( \Sigma_v^{-1} \otimes I_T \right) [z - (I_m \otimes Z) c] \right\}.
\]
The Likelihood

After some tedious algebra we get:

$$
\log \mathcal{L}(c, \Sigma_v) \propto |\Sigma_v|^{\frac{k}{2}} \exp \left\{ -\frac{1}{2} (c - c_{OLS})' \left( \Sigma_v^{-1} \otimes X'X \right) (c - c_{OLS}) \right\} \\
\times |\Sigma_v|^{-\frac{1}{2}(T-k)} \exp \left\{ -\frac{1}{2} \text{tr} \left[ \left( \Sigma_v^{-\frac{1}{2}} \otimes I_T \right) z - \left( \Sigma_v^{-\frac{1}{2}} \otimes X \right) c_{OLS} \right]' \right\} \\
\times \left[ \left( \Sigma_v^{-\frac{1}{2}} \otimes I_T \right) z - \left( \Sigma_v^{-\frac{1}{2}} \otimes X \right) c_{OLS} \right]
$$
Therefore,

\[
\log \mathcal{L}(c, \Sigma_v) \propto N \left( c | c_{OLS}, \Sigma_v, X, z \right) \\
\times W \left( \Sigma_v^{-1} | z, X, c_{OLS}, T - k - m - 1 \right).
\]

The likelihood is proportional to the product of a Gaussian distribution with a Wishart distribution.
Choosing the Priors

- Four main options:
  1. Gaussian for \( c \) with \( \Sigma_v \) fixed;
  2. diffuse prior for \( c \) and \( \Sigma_v \);
  3. Gaussian for \( c \) and diffuse for \( \Sigma_v \);
  4. Gaussian for \( c \) and Wishart for \( \Sigma_v \).

- In the first case let’s consider \( c \sim N(\overline{c}, \overline{\Sigma}_c) \).

- The posterior for \( c \) is given by:

\[
g(c|z) \propto \exp \left[ (c - \tilde{c})' \tilde{\Sigma}_c^{-1} (c - \tilde{c}) \right],
\]

where:

\[
\tilde{c} = \left[ \overline{\Sigma}_c^{-1} + \left( \Sigma_v^{-1} \otimes X'X \right) \right]^{-1} \left[ \overline{\Sigma}_c^{-1} \overline{c} + \left( \Sigma_v^{-1} \otimes X'X \right) z \right]
\]

\[
\tilde{\Sigma}_c = \left[ \overline{\Sigma}_c^{-1} + \left( \Sigma_v^{-1} \otimes X'X \right) \right]^{-1}.
\]
The Minnesota Prior

- For equation of the VAR the elements of $\tau$ related to the first lag of each dependent variable have unit mean. All the others have zero mean.
- $\Sigma_c$ is diagonal with $\sigma_{ij,\ell}$, corresponding to the lag $\ell$ of variable $j$ in equation $i$, given by:

$$
\sigma_{ij,\ell} = \begin{cases} 
\frac{\phi_0}{h(\ell)} & \text{se } i = j, \forall \ell, \\
\phi_0 \times \frac{\phi_1}{h(\ell)} \times \left( \frac{\sigma_i}{\sigma_j} \right)^2 & \text{se } i \neq j, j \text{ endogenous}, \forall \ell \\
\phi_0 \times \phi_2 & \text{se } i \neq j, j \text{ exogenous.}
\end{cases}
$$

The prior SSVS is a mixture of two Gaussian distributions:

\[ c_j | \gamma_j \sim (1 - \gamma_j)N(0, \kappa_{0j}^2) + \gamma_j N(0, \kappa_{1j}^2), \]

where:

- \( \gamma_j \) is a binary variable;
- \( \gamma_j = 1 \Rightarrow c_j | \gamma_j \sim N(0, \kappa_{1j}^2) \)
- \( \gamma_j = 0 \Rightarrow c_j | \gamma_j \sim N(0, \kappa_{0j}^2) \)

The prior SSVS is an hierarchic prior as \( \gamma_j \) não is not known.
Therefore,

\[ c | \gamma \sim N(0, DD), \]

where \( D \) is a diagonal matrix with the element \((j, j)\) given by:

\[ d_j = \begin{cases} \kappa_{0j} \text{ se } \gamma_j = 0, \\ \kappa_{1j} \text{ se } \gamma_j = 1. \end{cases} \]

\[ \kappa_{0j} = c_0 \sqrt{\text{var}(c_j)} \quad \text{e} \quad \kappa_{1j} = c_1 \sqrt{\text{var}(c_j)}, \quad c_0 << c_1. \quad \text{var}(c_j) \text{ is the standard deviation of the OLS estimator}. \]
A prior for $\gamma$:

$$\mathbb{P}(\gamma_j = 1) = q_j$$
$$\mathbb{P}(\gamma_j = 0) = 1 - q_j$$
The Structural VAR

- Write a structural VAR as:
  
  \[ ZB - XC = V. \]

- Make \( W = (Z, -X) \), \( A = (B, C)' \) and \( a = \text{vec}(A) \).

- The likelihood is:
  
  \[
  \log \mathcal{L}(A|W) \propto |B|^T \exp \left[-0.5a'(I \otimes W'W)a \right].
  \]
Let’s consider the following prior:

\[
\pi(a) = \pi(a_B)\pi(a_C|a_B).
\]

and

\[
\pi(a_C|a_B) = \mathcal{N}[h(B), \Sigma(B)].
\]

Hence,

\[
\pi(a|W) \propto |B|^T \exp \left[ -0.5 a'(I \otimes WW') a \right] |\Sigma(B)|^{-0.5} \\
\times \exp \left\{ -0.5 [a_C - h(B)]' \Sigma(B)^{-1} [a_C - h(B)] \right\} \pi(a_B).
\]