Roots from Trees:
A Machine Learning Approach to Unit Root Detection

Gary Cornwall
Bureau of Economic Analysis

Jeff Chen
Bennett Institute for Public Policy
University of Cambridge

Beau Sauley
Murray State University

October 20, 2021

The views expressed here are those of the authors and do not represent those of the U.S. Bureau of Economic Analysis or the U.S. Department of Commerce.
The [unresolved] presence of a unit root in time series can produce “nonsense regressions” (Granger & Newbold, 1974).

Dozens of tests have since been developed to detect unit roots under a variety of settings (e.g., panel data, structural breaks, etc.).

Tests don’t always agree but each tells you something about the series, how does a practitioner weigh the evidence?
1. Can we draw a link between a single unit root test and a “weak learner”.
2. If a test can be considered a weak learner, can we exploit between test variation using modern machine learning algorithms to better identify unit root processes?
Quick Answers

1. Can we draw a link between a single unit root test and a “weak learner”. Yes, in fact these are equivalent in both single and two-tailed tests for some $\alpha = \alpha'$.

2. If a test can be considered a weak learner, can we exploit between test variation using modern machine learning algorithms to better identify unit root processes? Yes, since we know how to aggregate weak base learners and create more powerful ensemble prediction methods we can use tools such as random forests and gradient boosting to improve unit root test accuracy.
Why Unit Roots?

- The unit root problem is a difficult time series econometrics problem which has produced nearly five decades of research and many different test statistics.
- The test for unit roots is important because failing to identify a unit root can invalidate all subsequent inferences (Granger & Newbold, 1974).
- Co-integrated relationships between series means you can’t just assume everything has a unit root (Granger, 1981; Engle and Granger, 1987).
- The difficulty comes from differentiating unit roots from near unit roots, as a result these test statistics have low power (Ng & Perron, 2001).
What is a Unit Root?

Let $y_t$ be an autoregressive time series generated such that,

$$y_t = \phi y_{t-1} + \epsilon_t, \quad t = (1, \ldots, T)$$

- We assume $\epsilon_t \sim N(0, \sigma^2) \forall t$ and that $\sigma_1^2 = \ldots = \sigma_T^2$.
- We can write this as $(1 - \phi L)y_t = \epsilon_t$ such that $Ly_t = y_{t-1}$.
- $(1 - \phi L)$ has a root of $1/\phi$ and if $|\phi| < 1$ then $y_t$ is considered stationary.
- Tests are often using an $H_0 : \phi = 1$ and $H_1 : |\phi| < 1$ structure (e.g. Augmented Dickey Fuller test).
How do we test for a Unit Root?

▶ The Dickey-Fuller test statistic:

$$\hat{\tau} = (\hat{\phi} - 1)S_e^{-1}\left(\sum_{t=2}^{N} Y_{t-1}^2\right)^{1/2},$$

$$S_e^{-1} = (n - 2)^{-1}\sum_{t=2}^{N}(Y_t - \hat{\phi}Y_{t-1})^2,$$

with limiting distribution outlined in Dickey & Fuller (1979).

▶ Calculated on first difference of $Y$ with $H_0: \phi = 1$ and $H_1: |\phi| < 1$. 
How do we test for a Unit Root?

Different assumed DGPs result in different null distributions and decision thresholds:

\[ y_t = \lambda + \phi y_{t-1} + \delta t + \epsilon_t \rightarrow x_\alpha = -3.45 | \alpha = 0.05 \]
\[ y_t = \lambda + \phi y_{t-1} + \epsilon_t \rightarrow x_\alpha = -2.89 | \alpha = 0.05 \]
\[ y_t = \phi y_{t-1} + \epsilon_t \rightarrow x_\alpha = -1.95 | \alpha = 0.05 \]

\[ \quad \]

Not everyone even agrees on the decision thresholds for the same test (e.g., Banerjee, et al., 1993 versus Hamilton, 1994 versus MacKinnon, 2010!)

---

1 Not everyone even agrees on the decision thresholds for the same test (e.g., Banerjee, et al., 1993 versus Hamilton, 1994 versus MacKinnon, 2010!)

---

8 / 21
How do we test for a Unit Root?

- Different assumed DGPs result in different null distributions and decision thresholds:¹

\[
y_t = \lambda + \phi y_{t-1} + \delta t + \epsilon_t \rightarrow x_\alpha = -3.45 | \alpha = 0.05
\]

\[
y_t = \lambda + \phi y_{t-1} + \epsilon_t \rightarrow x_\alpha = -2.89 | \alpha = 0.05
\]

\[
y_t = \phi y_{t-1} + \epsilon_t \rightarrow x_\alpha = -1.95 | \alpha = 0.05
\]

- Choice of DGP opens up an additional error path beyond Type I and Type II errors.

¹Not everyone even agrees on the decision thresholds for the same test (e.g., Banerjee, et al., 1993 versus Hamilton, 1994 versus MacKinnon, 2010!)
How do we test for a Unit Root?

▶ Different assumed DGPs result in different null distributions and decision thresholds:\(^1\)

\[
y_t = \lambda + \phi y_{t-1} + \delta t + \epsilon_t \rightarrow x_\alpha = -3.45 |\alpha = 0.05
\]

\[
y_t = \lambda + \phi y_{t-1} + \epsilon_t \rightarrow x_\alpha = -2.89 |\alpha = 0.05
\]

\[
y_t = \phi y_{t-1} + \epsilon_t \rightarrow x_\alpha = -1.95 |\alpha = 0.05
\]

▶ Choice of DGP opens up an additional error path beyond Type I and Type II errors.

▶ There are many tests that are similarly structured, e.g., ADF (Dickey and Fuller, 1981), PP (Phillips and Perron, 1988), KPSS (Kwiatkowski et al., 1992), PGFF (Pantula et al., 1994), Breit (Breitung, 2002; Breitung and Taylor, 2003), ERS (Elliot et al., 1996), URSP (Schmidt and Phillips, 1992), and URZA (Zivot and Andrews, 2002).

\(^1\)Not everyone even agrees on the decision thresholds for the same test (e.g., Banerjee, et al., 1993 versus Hamilton, 1994 versus MacKinnon, 2010!)
In both cases a decision is being made over a shared support of $\mathcal{X}$.

For some $\alpha = \alpha'$ it must be the case that $x_0 = x_\alpha$.

\[
(h(x) \equiv g(x)) |_{\alpha = \alpha'}
\]

where $h(x)$ is the decision stump and $g(x)$ is the Unit Root test.

$x_0 \approx -1.03$ which means $\alpha = \alpha' \approx 0.273$. 

Roots are Stumps
A Simple Procedure for Composite Test Construction

1. Simulate a balanced training, validation, and test set containing representative cases of the null and alternative hypotheses
2. Derive transmitters from one or multiple test statistics and attributes of the time series
3. Train a set of supervised classifiers, then select the model that fairs the best in cross-validation
4. Finally, conditional upon some desired Type I error rate, $\alpha$, or error cost ratio, $c(e_2)/c(e_1)$, return a class prediction for the series in question.
Simulate a balanced, representative data set

For any hypothesis test we can write down a DGP which will satisfy the null, e.g. unit roots.

1. Generate 500,000 time series with 350,000 for training, 75,000 for validation, and 75,000 for testing.
2. A series will contain a unit root, that is \( \phi = 1 \) with probability 0.50 and \( \phi \in \{0.9000, 0.9999\} \) otherwise.
3. Series will be uniformly distributed over the three unit root DGPs mentioned earlier.
4. All noise is Gaussian white noise.
## What are the features?

<table>
<thead>
<tr>
<th>UR Tests</th>
<th>Level and First Difference</th>
<th>STL Decomposed Series</th>
<th>Miscellaneous</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADF</td>
<td>Skewness</td>
<td>TNN Test</td>
<td>Length</td>
</tr>
<tr>
<td>PP</td>
<td>Kurtosis</td>
<td>Skewness</td>
<td>Frequency</td>
</tr>
<tr>
<td>PGFF</td>
<td>Box Statistic</td>
<td>Kurtosis</td>
<td>var(Δy)/var(y)</td>
</tr>
<tr>
<td>KPSS</td>
<td>Lyapunov Exponent</td>
<td>Box Statistic</td>
<td></td>
</tr>
<tr>
<td>ERS (d &amp; p)</td>
<td>TNN Test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>URSP</td>
<td>Hurst Exponent</td>
<td></td>
<td></td>
</tr>
<tr>
<td>URZA</td>
<td>Strength of Trend</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Breit</td>
<td>Strength of Seasonality</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

While we generate the data from one of the three possible “cases” outlined in the literature all test statistics are calculated on the most parsimonious DGP assumption possible, e.g. no drift or trend for the ADF.
Is there variation in our features?
Power Curves

Full Sample

DGP #1

DGP #2

DGP #3

Legend: ADF, ERS-d, GB, PP
Empirical Example

- Revisit 14 macro indicators from Nelson & Plosser (1982) which serve as a common benchmarking data set for unit root studies.
- Original paper indicated that, of the 14 series, only Unemployment Rate was stationary.
- We find that, depending on the desired Type I error rate, between 11 indicators ($\alpha = 0.10$) and 2 indicators ($\alpha = 0.01$) can be considered stationary.
```r
15 out01 <- ml_test(nelson_plosser_data, p.value = .01)
16 out05 <- ml_test(nelson_plosser_data, p.value = .05)
17 out10 <- ml_test(nelson_plosser_data, p.value = .10)
```
### thresholds_display

<table>
<thead>
<tr>
<th>method</th>
<th>pvalue</th>
<th>threshold</th>
<th>sensitivity</th>
<th>specificity</th>
<th>tp</th>
<th>tn</th>
<th>fp</th>
<th>fn</th>
<th>accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>xgbTree</td>
<td>0.51944008</td>
<td>0.9598049</td>
<td>0.9261502</td>
<td>35847</td>
<td>34884</td>
<td>2780</td>
<td>1509</td>
<td>0.9428133</td>
<td></td>
</tr>
<tr>
<td>xgbTree</td>
<td>0.66509449</td>
<td>0.9765232</td>
<td>0.9001016</td>
<td>36479</td>
<td>33880</td>
<td>3764</td>
<td>877</td>
<td>0.9381200</td>
<td></td>
</tr>
<tr>
<td>xgbTree</td>
<td>0.05</td>
<td>0.33739650</td>
<td>0.9249652</td>
<td>0.9500953</td>
<td>34553</td>
<td>35762</td>
<td>1882</td>
<td>2803</td>
<td>0.9375333</td>
</tr>
<tr>
<td>xgbTree</td>
<td>0.01</td>
<td>0.04205808</td>
<td>0.6550755</td>
<td>0.9900117</td>
<td>24471</td>
<td>37288</td>
<td>376</td>
<td>12885</td>
<td>0.8231867</td>
</tr>
</tbody>
</table>

### results

<table>
<thead>
<tr>
<th>series_no</th>
<th>pvalue_requested</th>
<th>pvalue Returned</th>
<th>type_of_results</th>
<th>score_xgbTree</th>
<th>threshold_xgbTree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01 using exact p-value threshold</td>
<td>0.0540929197</td>
<td>0.04205808</td>
</tr>
<tr>
<td>2</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01 using exact p-value threshold</td>
<td>0.126739885</td>
<td>0.04205808</td>
</tr>
<tr>
<td>3</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01 using exact p-value threshold</td>
<td>0.005586997</td>
<td>0.04205808</td>
</tr>
<tr>
<td>4</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01 using exact p-value threshold</td>
<td>0.931120264</td>
<td>0.04205808</td>
</tr>
<tr>
<td>5</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01 using exact p-value threshold</td>
<td>0.908699803</td>
<td>0.04205808</td>
</tr>
<tr>
<td>6</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01 using exact p-value threshold</td>
<td>0.216773159</td>
<td>0.04205808</td>
</tr>
<tr>
<td>7</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01 using exact p-value threshold</td>
<td>0.010059532</td>
<td>0.04205808</td>
</tr>
<tr>
<td>8</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01 using exact p-value threshold</td>
<td>0.027488123</td>
<td>0.04205808</td>
</tr>
<tr>
<td>9</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01 using exact p-value threshold</td>
<td>0.171301584</td>
<td>0.04205808</td>
</tr>
<tr>
<td>10</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01 using exact p-value threshold</td>
<td>0.115318856</td>
<td>0.04205808</td>
</tr>
<tr>
<td>11</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01 using exact p-value threshold</td>
<td>0.892714200</td>
<td>0.04205808</td>
</tr>
<tr>
<td>12</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01 using exact p-value threshold</td>
<td>0.894628441</td>
<td>0.04205808</td>
</tr>
<tr>
<td>13</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01 using exact p-value threshold</td>
<td>0.246658752</td>
<td>0.04205808</td>
</tr>
<tr>
<td>14</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01 using exact p-value threshold</td>
<td>0.827045248</td>
<td>0.04205808</td>
</tr>
</tbody>
</table>
Comparison with previous literature...
Unit Root tests are weak learners.

We can aggregate weak learners using gradient boosting to form a pseudo-composite test for unit roots.

This is [pessimistically] 20 percentage points more accurate and 37 percentage points more powerful than a traditional unit root test.
Thank you!
### Main Results Table

**Table: Main Results**

<table>
<thead>
<tr>
<th>Method</th>
<th>ACC</th>
<th>SEN</th>
<th>SPE</th>
<th>PPV</th>
<th>NPV</th>
<th>$F^1$</th>
<th>MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>GB $</td>
<td>\alpha = 0.100$</td>
<td>0.937</td>
<td>0.951</td>
<td>0.923</td>
<td>0.925</td>
<td>0.950</td>
<td>0.938</td>
</tr>
<tr>
<td>GB $</td>
<td>\alpha = 0.074^*</td>
<td>$</td>
<td>0.941</td>
<td>0.934</td>
<td>0.948</td>
<td>0.947</td>
<td>0.935</td>
</tr>
<tr>
<td>GB $</td>
<td>\alpha = 0.050$</td>
<td>0.938</td>
<td>0.900</td>
<td>0.976</td>
<td>0.973</td>
<td>0.908</td>
<td>0.935</td>
</tr>
<tr>
<td>GB $</td>
<td>\alpha = 0.010$</td>
<td>0.892</td>
<td>0.785</td>
<td>0.998</td>
<td>0.998</td>
<td>0.823</td>
<td>0.879</td>
</tr>
<tr>
<td>ADF</td>
<td>0.763</td>
<td>0.546</td>
<td>0.980</td>
<td>0.964</td>
<td>0.684</td>
<td>0.697</td>
<td>0.583</td>
</tr>
<tr>
<td>PP</td>
<td>0.744</td>
<td>0.512</td>
<td>0.975</td>
<td>0.953</td>
<td>0.667</td>
<td>0.666</td>
<td>0.549</td>
</tr>
<tr>
<td>KPSS</td>
<td>0.614</td>
<td>0.250</td>
<td>0.977</td>
<td>0.916</td>
<td>0.567</td>
<td>0.393</td>
<td>0.331</td>
</tr>
<tr>
<td>PGFF</td>
<td>0.745</td>
<td>0.499</td>
<td>0.989</td>
<td>0.978</td>
<td>0.665</td>
<td>0.661</td>
<td>0.560</td>
</tr>
<tr>
<td>BREIT</td>
<td>0.672</td>
<td>0.361</td>
<td>0.981</td>
<td>0.951</td>
<td>0.607</td>
<td>0.524</td>
<td>0.437</td>
</tr>
<tr>
<td>ERSd</td>
<td>0.762</td>
<td>0.545</td>
<td>0.979</td>
<td>0.963</td>
<td>0.683</td>
<td>0.696</td>
<td>0.582</td>
</tr>
<tr>
<td>ERSp</td>
<td>0.770</td>
<td>0.564</td>
<td>0.976</td>
<td>0.958</td>
<td>0.692</td>
<td>0.710</td>
<td>0.592</td>
</tr>
<tr>
<td>URZA</td>
<td>0.635</td>
<td>0.309</td>
<td>0.959</td>
<td>0.883</td>
<td>0.582</td>
<td>0.458</td>
<td>0.354</td>
</tr>
<tr>
<td>URSP</td>
<td>0.727</td>
<td>0.552</td>
<td>0.903</td>
<td>0.850</td>
<td>0.669</td>
<td>0.669</td>
<td>0.485</td>
</tr>
</tbody>
</table>