Clustering of Time Series

Raja Velu

JPMC Faculty Fellow
Syracuse University

April 4, 2012
Clustering of Time Series

1. Why?
   - Three Applications

2. How?
   - An Illustration
   - Review of Methods
   - Some New Ideas: Feature Weights, Lower Rank Approximation and Tracking Changes
   - Illustration Revisited

3. Large Scale Implementation
   - Scope & Limitations
Statistical Arbitrage: Pairs Trading

- Identify two stocks with historical prices moving together.
- When spread widens, short the winner and buy the loser.
- If history repeats itself, prices will converge and arbitrageur will profit.

Gatev, Goetzmann and Rouwenhorst, JFE, 2006
What is the Algorithm?

- Normalize stock prices over one year horizon.
- Select pairs with smallest normalized price difference*.
- If the normalized prices move beyond two standard deviations of historical difference, buy the cheap stock and sell the expensive one.
- Close out if the normalized prices cross; if the convergence does not happen within a specified period, liquidate the position.

*Euclidean distance
• Clustering of keywords in sponsored search.
• Advertisers buy keywords to bid; identify patterns from short sequences.
• Y! has 60m visitors per day; need a real time quick clustering algorithm.

Gluhovsky, Velu and Nagarajan, 2008, Patent Application, Y!
Basis Assets: Portfolio Analysis

- Opportunity set based on portfolios dominate that of individual assets
- Definition of Basis - representative set of portfolios
- Highly correlated assets are grouped together.

Ahn, Conrad and Dittman, 2009, RFS
Fig 1. Time Series Plot of Exchange Rates

Average Linkage ∼ Mixture Model
Why are the similarity values higher?
Clustering Approach

- Chronological Data
  - Non-stationary
    - Stationary
      - Data-Based
      - Feature-Based
        - Mixture Models
          - Or
          - Others
            - Variable Selection
            - # Clusters
              - Regularization Methods
            - Differencing, Co-integration, etc.
Static Data: Methodology

- Partitioning Methods
- Hierarchical Methods
- Density-Based Methods
- Grid-Based Methods
- Model-Based Methods
Raw Data Based Methods

- Distance measures are similar to Static Data
- Variance-Covariance matrices are too general
- No specific features are focused
Feature Based Distance Measures


Consider ARMA: $\phi_p(B)Y_t = \mu + \theta_v(B)a_t$

1. $d_{ij}^2 = \sum_{k=1}^{\alpha} (\Pi_{ik} - \Pi_{jk})^2$ Piccolo, 1990
2. $d_{ij}^2 = \sum_{k=1}^{\beta} (\Psi_{ik} - \Psi_{jk})^2$
3. $d_{ij}^2 = (\hat{\rho}_i - \hat{\rho}_j)'\Omega(\hat{\rho}_i - \hat{\rho}_j)$ Galeano & Pena, 2000

4. Kullbach-Leibler: $I(x, y) = tr(R_x R_y)^{-1} - \log \left(\frac{|R_x|}{|R_y|}\right) - n$

$R_x$ & $R_y$ are the $k \times k$ autocorrelation matrices

5. Chernoff’s Measure: $B_\alpha(x, y) = 
\frac{1}{2} \left[ \log \frac{\alpha R_x + (1 - \alpha)R_y}{|R_y|} - \alpha \log \frac{|R_x|}{|R_y|} \right]$

Focus mainly on the ACF/PACF and not on the level or variance ARMA representation is not unique.
Normalized Periodogram (Frequency Domain)

Let $P_x(w_j) = n^{-1} \left| \sum_{t=1}^{n} x_t e^{-itw_j} \right|^2$, $w_j = \frac{2\pi j}{n}$

$NP(w_j) = P(w_j/\hat{\gamma}_0) = 2 \left[ 1 + 2 \sum_{k=1}^{n-1} \hat{\rho}_k \cos(w_{jk}) \right]$  

$d_{NP}^2(x, y) = [NP_x(w_j) - NP_y(w_j)]^2$  

$= \frac{4}{n} d_{ACF}^2(x, y)$  

Caido, Crato & Pena, 2006
More Formal Methods (Feature Based)

\[ x_t \sim AR(k_1); \quad y_t \sim AR(k_2); \quad k = \text{Max}(k_1, k_2); \]
\[ x = w_x \Pi_x + a_x; \quad y = w_y \Pi_y + a_y; \quad Z = w \Pi + a; \]
\[ H_0 : \Pi_x = \Pi_y \iff R \Pi = 0 \]
\[ D = (R \hat{\Pi})' [R(w'(\hat{v})^{-1}w)^{-1}R']^{-1} R \hat{\Pi} \sim \chi^2_k \]

Use the p-values of this test to merge or demerge.
Maharaj, 2000
More Formal Methods (contd.)

For multivariate version ($m$ series),

$$D \sim \chi^2(m^2k)$$

Note: $H_0 : R\Pi = 0$, can be generalized to $R'$ unknown. If the series have same co-movements, the coefficient matrix $\Pi$ should reflect that. There is a close connection between clustering and co-movement idea in economics!
LR-Based Distance

Consider $AR(p): \phi_p(B)Y_t = \mu + a_t$, $a_t \sim N(0, \sigma^2)$
Let $\theta = (\mu, \phi_1, \ldots, \phi_p, \sigma^2)$; For two AR Models
$\sqrt{T}(\hat{\theta}_1 - \hat{\theta}_2)'G(\hat{\theta}_1 - \theta_2) \sim \chi^2_{p+2}$
Further Weigh $\theta$ by $W' \leftarrow$ Tuning Constants, $W'\theta$
Construct Distances $(W'\hat{\theta}_1 - W'\hat{\theta}_2)G^*(W\hat{\theta}_1 - W\hat{\theta}_2)$
Intuitive Ideas: $\theta = (\text{level, autocorrelation, variance})$ Features
$\sim$ Asymptotically independent/Combine them.

Two $AR(1): d^2 =$
\[
\left(\frac{\bar{y}_1 - \bar{y}_2}{s_p/\sqrt{T}}\right)^2 + \left(\hat{\rho}_1 - \hat{\rho}_2\right)^2 + \left(\frac{\hat{\sigma}_1^2/\hat{\sigma}_2^2}{(\kappa - 1)/\sqrt{T}}\right)^2 \sim \chi^2_3
\]
# Key Characteristics

**Table: Key Characteristics**

<table>
<thead>
<tr>
<th>variables</th>
<th>average</th>
<th>variance</th>
<th>Kurtosis</th>
<th>AutoCorr</th>
</tr>
</thead>
<tbody>
<tr>
<td>EUR</td>
<td>-0.0005</td>
<td>0.0001</td>
<td>1.78</td>
<td>0.28</td>
</tr>
<tr>
<td>USD</td>
<td>-0.0004</td>
<td>0.0001</td>
<td>4.23</td>
<td>0.28</td>
</tr>
<tr>
<td>GBP</td>
<td>0.0003</td>
<td>0.0002</td>
<td>3.62</td>
<td>0.29</td>
</tr>
<tr>
<td>CHF</td>
<td>-0.0013</td>
<td>0.0001</td>
<td>4.02</td>
<td>0.31</td>
</tr>
<tr>
<td>JPY</td>
<td>-0.0015</td>
<td>0.0002</td>
<td>2.30</td>
<td>0.30</td>
</tr>
<tr>
<td>CAD</td>
<td>-0.0009</td>
<td>0.0001</td>
<td>0.68</td>
<td>0.32</td>
</tr>
<tr>
<td>AUD</td>
<td>-0.0013</td>
<td>0.0002</td>
<td>11.91</td>
<td>0.32</td>
</tr>
<tr>
<td>CNY</td>
<td>-0.0011</td>
<td>0.0001</td>
<td>4.31</td>
<td>0.39</td>
</tr>
<tr>
<td>RUB</td>
<td>0.0000</td>
<td>0.0002</td>
<td>8.63</td>
<td>0.34</td>
</tr>
<tr>
<td>SAR</td>
<td>-0.0004</td>
<td>0.0001</td>
<td>4.35</td>
<td>0.37</td>
</tr>
<tr>
<td>BRL</td>
<td>-0.0010</td>
<td>0.0002</td>
<td>14.03</td>
<td>0.13</td>
</tr>
<tr>
<td>KRW</td>
<td>0.0002</td>
<td>0.0003</td>
<td>5.57</td>
<td>0.25</td>
</tr>
<tr>
<td>ZAR</td>
<td>-0.0004</td>
<td>0.0003</td>
<td>4.42</td>
<td>0.19</td>
</tr>
<tr>
<td>KPW</td>
<td>-0.0005</td>
<td>0.0001</td>
<td>4.29</td>
<td>0.38</td>
</tr>
<tr>
<td>ARS</td>
<td>0.0010</td>
<td>0.0001</td>
<td>4.10</td>
<td>0.36</td>
</tr>
</tbody>
</table>
Fig 5: LR Based Distance (Average Omitted)
Clustering Non-Stationary Series

- Dominant Feature: Non-stationarity component!
- Velu, Wichern and Reinsel (1987): Reduced-Rank model in Error Correction Form; provides a framework for testing, co-integration, co-movements.
Pairs Trading: Co-Integration / Clustering Application

- Stocks (by APT) with similar characteristics move together.
- If spread is large, trade.

Model:

\[
\begin{bmatrix}
 p_{1t} - p_{1t-1} \\
 p_{2t} - p_{2t-1}
\end{bmatrix}
= \begin{bmatrix}
 \alpha_1 \\
 \alpha_2
\end{bmatrix}
\begin{bmatrix}
 w_{t-1} - \mu_w \\
 \epsilon_{1t}
\end{bmatrix}
+ \begin{bmatrix}
 \epsilon_{1t} \\
 \epsilon_{2t}
\end{bmatrix}
\]

where \( w_t = p_{1t} - \gamma p_{2t} \) is mean reverting, also called spread.

- Long one share of stock 1, short \( \gamma \) shares of Stock 2, \( r_{p,t+i} = w_{t+i} - w_t \). Increment in the spread.
- Idea can be extended to multiple stocks that move together.
- What is the role of \( w_t \) in clustering?
Co-Integration Test

• VAR(2) Model: \( Y_t = \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + a_t \)
  \( w_t = Y_t - Y_{t-1} = \Phi_1^* Y_{t-1} + \Phi_2^* w_{t-1} + a_t \)

• Testing for unit roots \( \iff \) Testing for the rank \((\Phi_1^*)\)

• If \( \text{Rank}(\Phi_1^*) = r \), then we have \((m - r)\) unit roots.
  There are \( r \) linear combinations of \( Y_t \) that are unit-root stationary.

• Compute the \( \hat{\rho}^2 \), canonical correlations between \( W_t \) and \( Y_{t-1} \)
  and check if they are close to zero.

Johansen’s (1988) Test:
  \( -T \sum_{j=r+1}^{m} \log(1 - \hat{\rho}_j^2) \sim f(B_d(u)) \)
  \( B_d(u) \) is Brownian Motion; depends only on \( d = m - r \)!

• If \( \Phi_1^* \) is of rank \( r \), \( \Phi_1^* = \alpha \beta \)
  where \( \alpha \) is \( m \times r \) and \( \beta \) is \( r \times m \), then, \( X_t = \beta Y_t \) is unit root-stationary.
    • \( \beta \) - co-integrating vector.
Fig 6: Plot of Canonical Series

Johansen's Trace Test: Rank Three; Differs under regimes
Mixture Model Approach

- $Y_i \ (T \times 1) \ i^{th} \ series, \ i = 1, 2, \ldots, m$
- $f(Y, u/\theta, \pi) = f(u/\pi)f(Y/u, \theta)$
  $$= \left[\prod_{i=1}^{m} \prod_{k=1}^{K} \pi_{ik}^u\right] \left[\prod_{i=1}^{m} \prod_{k=1}^{K} f_k(Y_i/\theta)\right]^{u_{ik}}$$
- $K$ - clusters
- $u_{ik}$ - Probability that $i^{th}$ series belongs to the $k^{th}$ cluster
- $\pi_k$ - Prior probabilities

Use EM or MCMC

Computationally intensive / For Bayesian approach, need conformable priors

Xiong & Yeung (2004); Fruhwirth-Schnatter & Kaufmann (2008)
Clustering of Regression Models Method (CORM)

- Quin and Self (2006): Biometrics
- \( Y_i \) (\( T \times 1 \)) \( i^{th} \) series, \( i = 1, 2, \ldots, m \)
- \( X_i \) (\( T \times n \)) set of predictors; can include past values of \( Y_i \)

Model: \( E(Y_i) = X_i \beta_i \); \( Var(Y_i) = \sigma_i^2 I_T \)

\[ \beta_i \sim \sum_{k=1}^{K} u_{ik} N(\beta_k, \Sigma_k) \]

Estimate \( u_{ik}, \beta_k, \Sigma_k \) via GLS/EM algorithm.

- \( \beta_i \)'s are features.
Mixture Models: Main Issues

- Variable Selection: Adding more variables introduces noise
- How many clusters?
  (Friedman and Meulman (2004), JRSSB; Mangis et. al. (2009), Biometrics)
- For time-series data: clusters are not static!
Reduced-Rank Mixture Models

- Set-up: \( E(Y_i) = X_i \beta_i; \text{Var}(Y_i) = \sigma_i^2 I \) \( i = 1, 2, \ldots, m \)
- \( \beta_i \) (n x 1); n could be large
- Assume \( C = [\beta_1, \ldots, \beta_m]' \), m x n matrix is of lower rank
  \[ = A B_{mxr}r_{rxn} \]
- Thus \( E(Y_i) = X_i B' a_i \), where \( a_i \) is r x 1 \( \sim \sum_{k=1}^{K} u_{ik} N(a_k, \Sigma_k) \)
- \( B \) is common to all series; If \( r \leq 3 \), visualization methods are possible!
- Both variable selection and the choice of \( K \) can be achieved!
Fig 7: Scatterplot of a1 vs a2

a1 ~ general trend; a2 ~ ?
Some Additional Thoughts

- Time Domain: Decompose the series as trend, seasonality, and cycle; Group the series by these components.
- Large number of series: Exponential smoothing works well. Can we capture the structure via a single smoothing constant and base our groupings?
- Develop methods for clustering as a moving window. Clustering is not static in the time series context.
- How do we incorporate cost aspects of clustering?