Detecting Covariance Changes Using a Likelihood-Ratio Test

Cody Buntain

1 Overview

In Galeano and Peña's work on detecting covariances changes in multivariate data, they proposed two methods for calculating test statistics from which change points could be identified [1]. These methods model the given data as a vector autoregressive integrated moving average (vARIMA) popular in economics and financial market analysis, extracting the errors (or innovations) from this data, and applying these methods on this error data. The first such statistic, on which we focus here, uses a likelihood-ratio test (LRT) to compare two hypotheses: the null hypothesis H_n that the covariance of this error data is best characterized by a single covariance matrix Σ versus the alternative hypothesis H_a that, at some time point h, the data is best characterized by two separate covariances matrices Σ_1 before h and Σ_2 after h. The logarithm of a modified form of the ratio H_n/H_a then generates a test statistic LR_h that existing literature shows is governed by a chi-squared distribution with degrees of freedom proportional to the dimensionality k of the data. From simulations of this distribution, we can generate a critical value given some α against which to compare this test statistic to determine whether a change point actually exists at some time h.

2 Algorithm

Given some time-series data $\tilde{y}_t$ and confidence α, we use the following algorithm to identify points of change in covariance:

Function LRT($\tilde{y}_t, \alpha$) Algorithm by Galeano and Peña [1]

fit VARIMA(p, d', q) model to $\tilde{y}_t$;
compute residuals $\hat{e}_t$;

$k \leftarrow$ dimension($\tilde{y}_t$) ;
$d \leftarrow k(p+q+1) + \frac{k(k+1)}{2} + 1$; `/* minimum points needed */`
$n \leftarrow$ len($\tilde{y}_t$) ;
$df \leftarrow \frac{k(k+1)}{2}$; `/* degrees of freedom for` χ^2 `*/`
$C \leftarrow$ simulateChiSquareMax(df, α) ; `/* obtain the critical value */`

$LR \leftarrow$ zeros(n) ;
$S \leftarrow \frac{1}{n}\Sigma_{i=1}^{n} e_i \cdot e_i'$;
for $h \in [d, n-d-1]$ **do**
    $v \leftarrow h/n$;
    $S_1 \leftarrow \frac{1}{h}\Sigma_{i=1}^{h} e_i \cdot e_i'$;
    $S_2 \leftarrow \frac{1}{n-h}\Sigma_{i=h+1}^{n} e_i \cdot e_i'$;
    $LR[h] \leftarrow n \ln \frac{|S|}{|S_1|^v |S_2|^{1-v}}$;
end

$h_{max} \leftarrow argmax_h(LR)$;
$\Lambda_{max} \leftarrow LR[h_{max}]$;

changePoints $\leftarrow$ [ ] ;
if $\Lambda_{max} > C$ **then**
    changePoints += h_{max} ;
    $W \leftarrow$ transformation governing new data regime (see [1]);
    changePoints += apply LRT to $\hat{e}_t[0 : h_{max}]$;
    changePoints += apply LRT to $W \cdot \hat{e}_t[h_{max} + 1 : n]$;
end

return changePoints

2.1 Implementation Details

To implement LRT, we used Python and Scikit's statsmodels package for fitting data to VAR() models. One should note this restriction to VAR() models is a result of an existing constraint in the statsmodels package.

We also implemented a version of the LRT algorithm that does not rely on calculating the W transformation matrix. Rather than evaluating W, we leveraged statsmodels and its maximum likelihood estimation to fit the data to two new VAR() models for each regime. The above algorithm performs better than this secondary implementation because it obviates the need for separate rounds of maximum likelihood estimation for each level of recursion.

References

[1] P. Galeano and D. Peña. Covariance changes detection in multivariate time series. *Journal of Statistical Planning and Inference*, 137(1):194–211, Jan. 2007.