The Center for Artificial Intelligence in Business Analytics & Financial Technology

Robust Rolling PCA (R2-PCA)

Ali Hirsa, Federico Klinkert, Satyan Malhotra, Ryan Holmes

ask².ai

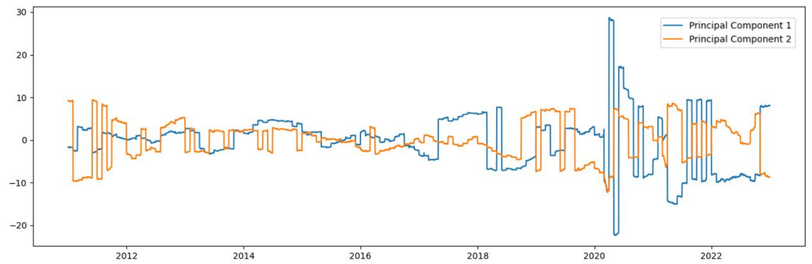
Columbia | Engineering
The Fu Foundation School of Engineering and Applied Science

Abstract

Principal Component Analysis (PCA) is an important methodology to reduce and extract meaningful signals from large data-sets. Financial markets introduce time and non stationarity aspects, where applying standard PCA methods may not give stable results. We propose robust rolling PCA (R2-PCA) that accommodates the additional aspects and mitigates commonly found obstacles including eigenvector sign flipping, and managing multiple dimensions of the data-set. This makes R2-PCA an ideal candidate for learning-based models.

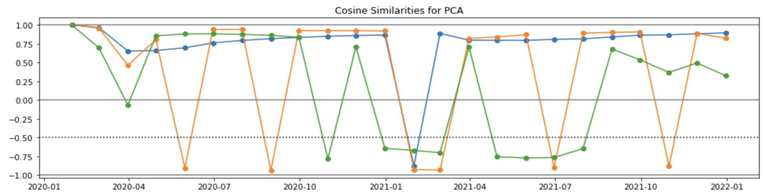
Eigenvector Sign Flipping

Eigenvector problems with the unit vector constraint typically yield more than one solution when decomposing a symmetric covariance matrices; effectively each eigenvector can be multiplied by -1 and still satisfy the constraints. Computer based PCA solvers choose the eigenvector in the direction of the data. If the direction of the data switches over time (a common occurrence in financial time series especially if the data is normalized), this can lead to the projected data jumping between time increments. This will drastically reduce the accuracy of reduced data to its original counterpart.



Cosine Similarity Score

To fix eigenvector flipping issue, the cosine similarity measure (dot product in most cases) is used to related the eigenvectors from different periods. Should an eigenvector flip between time increments, the similarity score will drop to a value below 0 and the eigenvector can be unflipped by multiplying it by -1.



R2-PCA Algorithm

Algorithm 1 R2-PCA Algorithm

Choose an number of Principal Components p
Choose a dataset X with dimensions (F, T, D) ▷ (Funds, Time, Features)
Choose a rolling window length W
Choose w as set of all time increments up to time t with $|w| = W$
Set time $t = 1$ and rolling window $w_t = \{t, t-1, ..., t-W+1\}$ ▷ If data exists for $t < 1$, else $w_t = \{t\}$
Compute the covariance matrix C_f for each element/asset in $f \in w_t$
Compute average covariance matrix $\bar{C} = \frac{1}{|F_t|} \sum_i C_i$
Eigendecompose $\bar{C} = P\Lambda P^T$ and extract eigenvectors $V_{w_t} = \{v_1, ...v_p\}_{w_t}$
for $t = 2, ..., T$ **do**
  Set rolling window $w_t = \{t, t-1, ..., t-W+1\}$
  Compute the covariance matrix C_f for each element/asset $f \in w_t$
  Compute average covariance matrix $\bar{C} = \frac{1}{|F_t|} \sum_i C_i$
  Eigendecompose $\bar{C} = P\Lambda P^T$ and extract eigenvectors $V_{w_t} = \{v_1, ..., v_p\}_{w_t}$
  for $i = 1, ..., p$ **do**
    Set $j = \text{argmax}(|v^i_{w_t} \cdot V_{w_{t-1}}|)$ ▷ Find eigenvector with highest absolute similarity score for ordering
    if $v^i_{w_t} \cdot v^j_{w_{t-1}} < 0$ **then** ▷ Eigenvectors from current and previous rolling windows with $||v|| = 1$
      Set $v^i_{w_t} = -v^i_{w_t}$ ▷ Sign Flip
    end if
  end for
  Reorder $V_{w_t} = \{v_1, ..., v_p\}_{w_t}$ from each argmax j result ▷ Data can be projected using V_{w_t} after this step
end for

Features of R2-PCA Model

- **Variable Number of Dataset Features:** R2-PCA can be used on datasets with a variable number of features over time by computing the cosine similarity of only the shared features between time periods

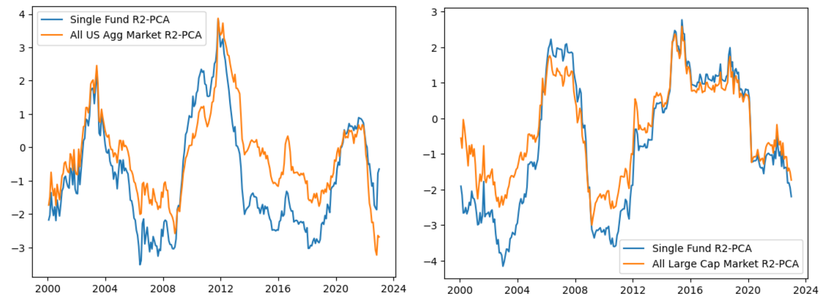
- **Reordering of Principal Components:** Cosine similarity can be computed across all eigenvectors in different time increments and compared. By taking the largest absolute value of the cosine similarity, we can reorder the principal components based on the magnitude of variation from the previous time increment

- **Incorporating More Dimensions:** Time series datasets with additional dimensions can utilize the R2-PCA algorithm by averaging the covariance matrix at each time increment. The datasets used for this analysis had the additional dimension of funds, thus the covariance at each time increment was computed by averaging the covariance for each fund.

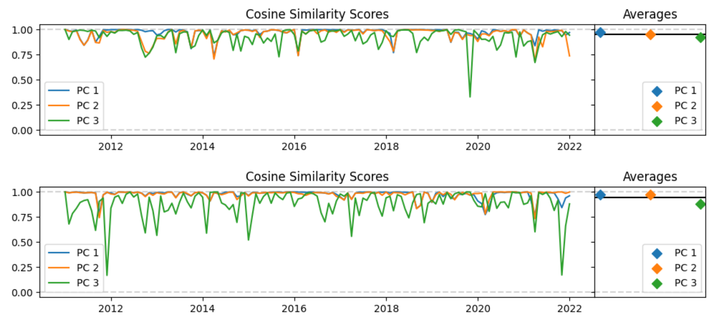
- **Covariance Stability:** The R2-PCA Model was specifically designed to handle financial time series datasets where the covariance matrix may significantly change due to the effects of market shocks and other regime dependent effects. A covariance instability back test was used to stress test R2-PCA and other PCA models. The results of these tests can be found in the SSRN Paper QR code and the Animations QR code link.

R2-PCA Results

The results above depict R2-PCA and PCA run on the entire dataset at one time. This gives a good idea as to what the shape of data should be if the future results of the market are known in the present.



R2-PCA Cosine Similarity Results



SSRN Paper

Animations