tôi muốn xác nhận viêc chuẩn hoá dữ liệu Cân chỉnh trung tâm tập dữ liệu (Mean Centering) sau khi xử lý ngoại lai với IRQ và phương pháp z-score với độ lệch $ \sigma_j $ của biến j có ảnh hưởng gì đến tính Ma trận hiệp phương sai không?

Để trả lời câu hỏi của bạn, chúng ta sẽ phân tích việc **chuẩn hóa dữ liệu** bằng **cân chỉnh trung tâm (mean centering)** sau khi xử lý ngoại lai với IQR, và việc áp dụng **phương pháp z-score** với độ lệch chuẩn σ_j của từng biến j, và tác động của chúng đến **ma trận hiệp phương sai** trong bối cảnh chuẩn bị dữ liệu cho PCA với tập dữ liệu `VNINDEX_iqr_filled_HSG_Adjust.csv`. Tôi sẽ giải thích chi tiết từng bước, bao gồm khái niệm, công thức, và ảnh hưởng cụ thể.

---

1. Các khái niệm cơ bản

a. Cân chỉnh trung tâm (Mean Centering)

- **Định nghĩa**: Trừ giá trị trung bình μ_j của mỗi biến (cột) j cho tất cả các giá trị trong cột đó, để trung bình của cột bằng 0.
- **Công thức**:

$$X'_{ij} = X_{ij} - \mu_j, \quad \text{với} \quad \mu_j = \frac{1}{n} \sum_{i=1}^{n} X_{ij}$$

  Trong đó:

  - X_{ij}: Giá trị gốc của biến j tại quan sát i.
  - X'_{ij}: Giá trị sau cân chỉnh trung tâm.
  - n: Số quan sát.

- **Mục đích**: Đảm bảo dữ liệu tập trung quanh gốc tọa độ (trung bình = 0), giúp PCA tính toán ma trận hiệp phương sai dựa trên sự biến động của dữ liệu, không bị ảnh hưởng bởi giá trị trung bình.

b. Chuẩn hóa z-score (Z-score Standardization)

- **Định nghĩa**: Sau khi cân chỉnh trung tâm, chia mỗi giá trị trong cột cho độ lệch chuẩn σ_j của cột đó, để mỗi biến có trung bình bằng 0 và độ lệch chuẩn bằng 1.
- **Công thức**:

$$Z_{ij} = \frac{X'_{ij}}{\sigma_j} = \frac{X_{ij} - \mu_j}{\sigma_j}$$

  Trong đó:

  - $\sigma_j = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (X_{ij} - \mu_j)^2}$: Độ lệch chuẩn của biến j.
  - Z_{ij}: Giá trị sau chuẩn hóa z-score.

- **Mục đích**: Đưa các biến về cùng thang đo (độ lệch chuẩn = 1), đảm bảo các biến có mức đóng góp tương đương vào PCA, đặc biệt khi các biến có đơn vị hoặc phạm vi giá trị khác nhau (ví dụ: `VNINDEX` ~ hàng trăm, `HSG_log` ~ đơn vị).

c. Ma trận hiệp phương sai

- **Định nghĩa**: Ma trận hiệp phương sai đo lường mức độ biến thiên chung giữa các cặp biến trong tập dữ liệu.
- **Công thức**:

$$C_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} (X_{ij} - \mu_j)(X_{ik} - \mu_k)$$

  Trong đó:

  - C_{jk}: Hiệp phương sai giữa biến j và biến k.
  - μ_j, μ_k: Trung bình của biến j và k.

- **Trong PCA**: Ma trận hiệp phương sai được sử dụng để tìm các thành phần chính (eigenvectors) và giá trị riêng (eigenvalues), đại diện cho hướng và độ lớn của phương sai dữ liệu.

d. Xử lý ngoại lai với IQR

- **Phương pháp**: Sử dụng khoảng tứ phân vị (IQR) để xác định và thay thế ngoại lai.
  - Tính: $Q1$ (tứ phân vị thứ nhất), $Q3$ (tứ phân vị thứ ba), $IQR = Q3 - Q1$.
  - Ngoại lai: Giá trị $< Q1 - 3 \cdot IQR$ hoặc $> Q3 + 3 \cdot IQR$ (hệ số 3.0 như bạn đã áp dụng cho `HSG_log`).
  - Thay thế: Ngoại lai được thay bằng giá trị biên ($Q1 - 3 \cdot IQR$ hoặc $Q3 + 3 \cdot IQR$) hoặc bằng log1p (như với `HSG_log`).
- **Tác động**: Giảm ảnh hưởng của các giá trị cực đại/cực tiểu, làm dữ liệu ít bị lệch (skewness), phù hợp hơn cho PCA.

---

2. Tác động của Mean Centering và Z-score đến Ma trận hiệp phương sai

a. Tác động của Mean Centering

- **Ảnh hưởng đến trung bình**:
  - Sau mean centering, mỗi biến j có trung bình $\mu_j' = 0$.
  - Điều này đảm bảo rằng ma trận hiệp phương sai được tính toán chỉ dựa trên sự biến động quanh gốc (0), không bị ảnh hưởng bởi giá trị trung bình ban đầu.
- **Ảnh hưởng đến ma trận hiệp phương sai**:
  - Nếu dữ liệu không được mean centered, trung bình $\mu_j \neq 0$ sẽ làm tăng giá trị hiệp phương sai, vì:

$$C_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} (X_{ij} - \mu_j)(X_{ik} - \mu_k)$$

Nếu $\mu_j \neq 0$, các giá trị X_{ij} sẽ bao gồm cả mức độ dịch chuyển trung bình, làm sai lệch hiệp phương sai.

- Sau mean centering ($X'_{ij} = X_{ij} - \mu_j$):

$$C'_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} X'_{ij} X'_{ik} = \frac{1}{n-1} \sum_{i=1}^{n} (X_{ij} - \mu_j)(X_{ik} - \mu_k)$$

Ma trận hiệp phương sai không thay đổi về bản chất, nhưng được tính chính xác hơn vì loại bỏ ảnh hưởng của trung bình.

- **Tác động sau xử lý IQR**:
  - Xử lý ngoại lai với IQR làm giảm các giá trị cực đại, giúp trung bình μ_j và phương sai ổn định hơn.
  - Mean centering sau IQR không làm thay đổi bản chất của ma trận hiệp phương sai, nhưng đảm bảo rằng hiệp phương sai phản ánh đúng sự biến động của dữ liệu đã được làm mịn (loại bỏ ngoại lai).

b. Tác động của Chuẩn hóa Z-score

- **Ảnh hưởng đến trung bình và độ lệch chuẩn**:
  - Sau z-score, mỗi biến có:
    - Trung bình: $\mu_j = 0$.
    - Độ lệch chuẩn: $\sigma_j = 1$.
  - Điều này đưa tất cả các biến về cùng thang đo, bất kể đơn vị hoặc phạm vi ban đầu (ví dụ: `VNINDEX` ~ hàng trăm, `HSG_log` ~ đơn vị).
- **Ảnh hưởng đến ma trận hiệp phương sai**:
  - Ma trận hiệp phương sai của dữ liệu z-score là **ma trận tương quan** (correlation matrix), vì:

$$C_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} Z_{ij} Z_{ik} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{X_{ij} - \mu_j}{\sigma_j}\right)\left(\frac{X_{ik} - \mu_k}{\sigma_k}\right)$$

$$C_{jk} = \frac{\text{Cov}(X_j, X_k)}{\sigma_j \sigma_k} = \varrho_{jk}$$

Trong đó ϱ_{jk} là hệ số tương quan Pearson giữa biến j và k.

- **So sánh**:
  - Nếu chỉ mean centering: Ma trận hiệp phương sai giữ nguyên phương sai và hiệp phương sai gốc, phản ánh độ lớn tuyệt đối của biến động.

    - Nếu áp dụng z-score: Ma trận hiệp phương sai trở thành ma trận tương quan, chỉ phản ánh mức độ tương quan tuyến tính giữa các biến, không phụ thuộc vào thang đo.
- **Tác động sau xử lý IQR**:
    - IQR làm giảm ảnh hưởng ngoại lai, giúp σ_j (độ lệch chuẩn) ổn định hơn, tránh bị thổi phồng bởi các giá trị cực đại.
    - Chuẩn hóa z-score sau IQR làm cho ma trận tương quan đáng tin cậy hơn, vì dữ liệu đã được làm mịn và các biến có thang đo thống nhất.

c. So sánh tác động đến PCA

- **Chỉ Mean Centering**:
    - Ma trận hiệp phương sai giữ nguyên phương sai của từng biến.
    - Các biến có thang đo lớn (ví dụ: `VNINDEX`) sẽ chi phối các thành phần chính, vì PCA ưu tiên hướng có phương sai lớn.
    - Phù hợp khi các biến có cùng đơn vị hoặc thang đo tương tự, hoặc khi bạn muốn giữ nguyên mức độ biến động tuyệt đối.
- **Mean Centering + Z-score**:
    - Ma trận hiệp phương sai trở thành ma trận tương quan.
    - Tất cả biến có đóng góp đồng đều vào PCA (do $\sigma_j = 1$), phù hợp khi các biến có đơn vị/thang đo khác nhau (như trong tập `VNINDEX_iqr_filled_HSG_Adjust.csv`).
    - Các thành phần chính phản ánh mức độ tương quan tuyến tính giữa các biến, thay vì phương sai tuyệt đối.

---

3. Ảnh hưởng cụ thể với tập dữ liệu `VNINDEX_iqr_filled_HSG_Adjust.csv`

a. Đặc điểm tập dữ liệu

- **Cột**: 36 cột (`VNINDEX`, `BVH`, `CSM`, ..., `CONGNGHE`), đều là số.
- **Xử lý ngoại lai**: Đã áp dụng IQR với hệ số 3.0, thay thế ngoại lai bằng giá trị biên hoặc log1p (cho `HSG_log`).
- **Dữ liệu thiếu**: Đã được nội suy tuyến tính.
- **Thang đo**: Các cột có thang đo khác nhau:
    - `VNINDEX`: ~300-700.
    - `HSG_log`: ~0.9-3.4 (sau log1p).
    - `KDC`, `VNM`, ...: Thang đo khác nhau (hàng chục đến hàng trăm).

b. Tác động của Mean Centering

- **Trên dữ liệu**:
    - Mỗi cột (`VNINDEX`, `HSG_log`, ...) được trừ trung bình của nó, ví dụ:
        - Nếu $\mu_{\text{VNINDEX}} \approx 475$, thì $VNINDEX' = VNINDEX - 475$.

    - Nếu $\mu_{\text{HSG\_log}} \approx 1.8$, thì $HSG_log' = HSG_log - 1.8$.
  - Trung bình của mỗi cột trở thành 0.
- **Trên ma trận hiệp phương sai**:
  - Ma trận hiệp phương sai phản ánh đúng sự biến động của dữ liệu sau khi loại bỏ trung bình.
  - Phương sai của các cột có thang đo lớn (như `VNINDEX`) sẽ lớn hơn, dẫn đến việc chúng chi phối các thành phần chính trong PCA.
  - Ví dụ: Hiệp phương sai giữa `VNINDEX` và `VNM`:

$$C_{\text{VNINDEX,VNM}} = \frac{1}{n-1} \sum (VNINDEX_i - \mu_{\text{VNINDEX}})(VNM_i - \mu_{\text{VNM}})$$

  - **Sau IQR**: Ngoại lai đã được kiểm soát, nên trung bình và hiệp phương sai ổn định hơn, tránh bị lệch bởi các giá trị cực.

c. Tác động của Z-score

- **Trên dữ liệu**:
  - Sau mean centering, mỗi cột được chia cho σ_j:
    - Nếu $\sigma_{\text{VNINDEX}} \approx 100$, thì $Z_{\text{VNINDEX}} = \frac{VNINDEX - \mu_{\text{VNINDEX}}}{100}$.
    - Nếu $\sigma_{\text{HSG\_log}} \approx 0.5$, thì $Z_{\text{HSG\_log}} = \frac{HSG_log - \mu_{\text{HSG\_log}}}{0.5}$.
  - Kết quả: Mỗi cột có $\mu_j = 0$, $\sigma_j = 1$.
- **Trên ma trận hiệp phương sai**:
  - Ma trận hiệp phương sai trở thành ma trận tương quan:

$$C'_{jk} = \varrho_{jk} = \frac{\text{Cov}(X_j, X_k)}{\sigma_j \sigma_k}$$

  - Các giá trị trong ma trận nằm trong $[-1, 1]$, phản ánh mức độ tương quan tuyến tính.
  - Các biến như `VNINDEX` và `HSG_log` có đóng góp đồng đều vào PCA, bất kể thang đo ban đầu.
  - **Sau IQR**: Độ lệch chuẩn σ_j ổn định hơn do ngoại lai đã được xử lý, nên ma trận tương quan đáng tin cậy hơn.

d. Tác động đến PCA

- **Chỉ Mean Centering**:
  - Các cột như `VNINDEX` (phương sai lớn) sẽ chi phối các thành phần chính đầu tiên.
  - Phù hợp nếu bạn muốn giữ nguyên phương sai gốc của các biến (ví dụ: khi tất cả biến đều quan trọng về mặt kinh tế và thang đo tự nhiên có ý nghĩa).
  - Tuy nhiên, các cột có thang đo nhỏ (như `HSG_log`) có thể bị lu mờ.
- **Mean Centering + Z-score**:
  - Ma trận tương quan đảm bảo tất cả biến đóng góp đồng đều.

- Phù hợp với tập dữ liệu của bạn, vì các cột có thang đo khác nhau (`VNINDEX` , `HSG_log` , `KDC` , ...).
- Các thành phần chính phản ánh các mối quan hệ tuyến tính giữa các biến, thay vì bị chi phối bởi biến có phương sai lớn.

---

4. Kết luận

- **Mean Centering**:
  - **Ảnh hưởng**: Loại bỏ trung bình, đưa trung bình mỗi cột về 0, đảm bảo ma trận hiệp phương sai phản ánh đúng sự biến động của dữ liệu.
  - **Sau IQR**: Trung bình và hiệp phương sai ổn định hơn, vì ngoại lai đã được kiểm soát.
  - **Tác động đến PCA**: Giữ nguyên phương sai gốc, các biến có thang đo lớn chi phối thành phần chính.
- **Z-score Standardization**:
  - **Ảnh hưởng**: Đưa trung bình về 0 và độ lệch chuẩn về 1, biến ma trận hiệp phương sai thành ma trận tương quan.
  - **Sau IQR**: Độ lệch chuẩn ổn định, ma trận tương quan đáng tin cậy hơn.
  - **Tác động đến PCA**: Các biến đóng góp đồng đều, phù hợp khi các biến có thang đo khác nhau.
- **So sánh**:
  - Mean centering là bước bắt buộc cho PCA, đảm bảo ma trận hiệp phương sai chính xác.
  - Z-score là tùy chọn, nhưng được khuyến nghị khi các biến có thang đo khác nhau (như trong tập `VNINDEX_iqr_filled_HSG_Adjust.csv`).
  - IQR giúp cả hai phương pháp ổn định hơn bằng cách giảm ảnh hưởng của ngoại lai.

---

5. Đề xuất cho tập dữ liệu của bạn

- **Khuyến nghị**:
  - Vì các cột trong `VNINDEX_iqr_filled_HSG_Adjust.csv` có thang đo khác nhau (`VNINDEX` ~ hàng trăm, `HSG_log` ~ đơn vị), bạn nên áp dụng **cả mean centering và z-score** để:
    - Đảm bảo trung bình = 0 (mean centering).
    - Đưa các biến về cùng thang đo (z-score).
    - PCA sẽ sử dụng ma trận tương quan, giúp tất cả biến (`VNINDEX` , `HSG_log` , `KDC` , ...) đóng góp đồng đều.