ST2DA-I2 | 2025-2026

December 14, 2025

1 Introduction

2 Mathematical Framework

3 Results

4 Interpretation

5 Conclusion

Definition

Principal Component Analysis is a **linear transformation** that projects high-dimensional data onto a lower-dimensional subspace while **maximizing variance**.

Key Properties:

- Orthogonal transformation
- Variance maximization
- Dimensionality reduction
- Feature extraction

Applications:

- Data visualization
- Noise reduction
- Feature engineering
- Exploratory analysis

True Cost of Fast Fashion Dataset

Overview:

- $n = 3000$ observations
- $p = 12$ quantitative variables
- 5 brands: Shein, Zara, H&M, Forever 21, Uniqlo
- 10 countries, 2015-2024

Variable Categories:

- Production: Volume, Cycles
- Environmental: CO_2, Water, Waste
- Sustainability: Scores, Indices

Why PCA?

Variables have **different scales** (tonnes, USD, %) $\Rightarrow$ Use **correlation-based PCA**

Objective

Find direction $\mathbf{v}_1$ that maximizes variance of projected data:

$$\mathbf{v}_1 = \arg\max_{\|\mathbf{v}\|=1} \mathrm{Var}(\mathbf{X}\mathbf{v}) = \arg\max_{\|\mathbf{v}\|=1} \mathbf{v}^\top \mathbf{S}\mathbf{v}$$

Lagrangian formulation:

$$\mathcal{L}(\mathbf{v}, \lambda) = \mathbf{v}^\top \mathbf{S}\mathbf{v} - \lambda(\mathbf{v}^\top \mathbf{v} - 1)$$

First-order condition:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{v}} = 2\mathbf{S}\mathbf{v} - 2\lambda\mathbf{v} = 0 \quad \Rightarrow \quad \boxed{\mathbf{S}\mathbf{v} = \lambda\mathbf{v}}$$

$\Rightarrow$ **Eigenvalue problem!**

Theorem (Spectral Theorem)

For symmetric matrix $\mathbf{R} \in \mathbb{R}^{p \times p}$*:*

$$\mathbf{R} = \mathbf{V}\,\mathbf{V}^\top = \sum_{k=1}^{p} \lambda_k \mathbf{v}_k \mathbf{v}_k^\top$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$

Key Results:

- Eigenvectors $\mathbf{v}_k$ are **orthonormal**: $\mathbf{V}^\top\mathbf{V} = \mathbf{I}$
- Eigenvalues λ_k = **variance** of k-th principal component
- Total variance preserved: $\sum_{k=1}^{p} \lambda_k = \mathrm{tr}(\mathbf{R}) = p$

Covariance Matrix $\mathbf{S}$

$$s_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

Use when:

- Variables have same units
- Scale matters

Correlation Matrix $\mathbf{R}$

$$r_{jk} = \frac{s_{jk}}{s_j \cdot s_k}$$

Use when:

- Variables have different units
- Want equal contribution

Our Choice: Correlation Matrix

Variables have heterogeneous units (tonnes, USD, %, indices) $\Rightarrow$ **Standardize first!**

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

Factor Loadings

Correlation between variable X_j and component Z_k:

$$\ell_{jk} = v_{jk} \cdot \sqrt{\lambda_k} = \mathrm{Corr}(X_j, Z_k)$$

Quality of Representation (cos²):

$$\cos^2_{j,K} = \sum_{k=1}^{K} \ell_{jk}^2$$

$\rightarrow$ How well is variable j represented?

Variable Contribution:

$$\mathrm{CTR}_{jk} = v_{jk}^2$$

$\rightarrow$ How much does j contribute to PC_k?

How many components to retain?

1. **Kaiser Criterion** (1960):

$$\text{Retain PC}_k \text{ if } \lambda_k > 1$$

   Rationale: Component should explain more than one original variable

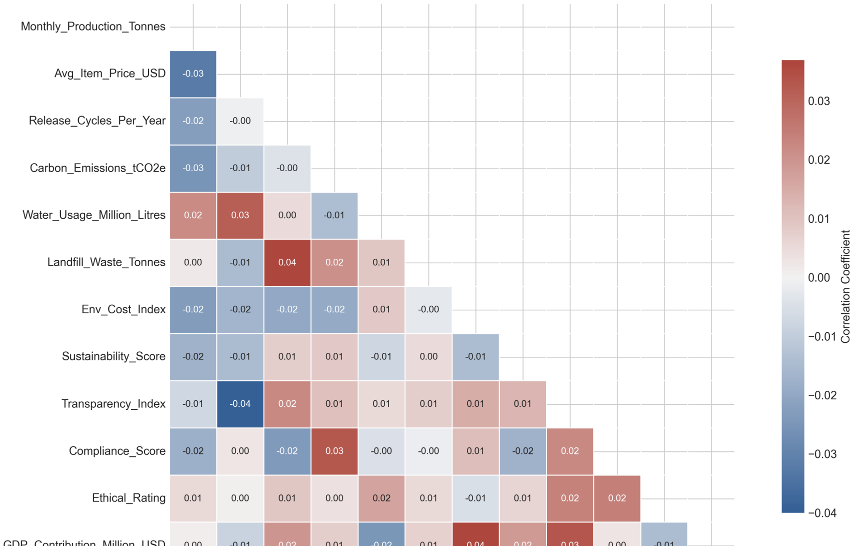
2. **Scree Test** (Cattell, 1966):

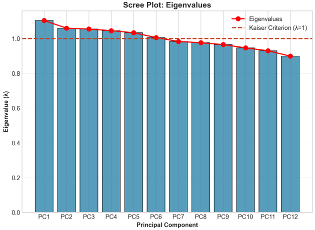
$$\text{Find “elbow” in } \{\lambda_1, \lambda_2, \ldots, \lambda_p\}$$

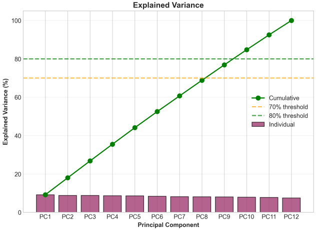
3. **Variance Threshold**:

$$\text{Retain } K \text{ s.t. } \sum_{k=1}^{K} \frac{\lambda_k}{p} \geq 0.70 \text{ or } 0.80$$

Correlation Matrix of Environmental and Sustainability Variables

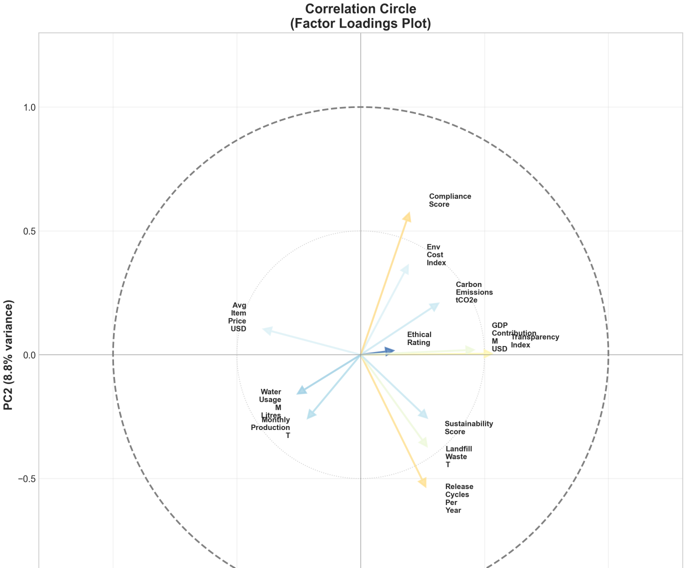






PC	λ	Cum.%
1	1.10	9.2
2	1.06	18.0
3	1.05	26.8
4	1.05	35.5
5	1.03	44.1
6	1.01	52.5
7	0.98	60.7
⋮	⋮	⋮

Kaiser: 6 components

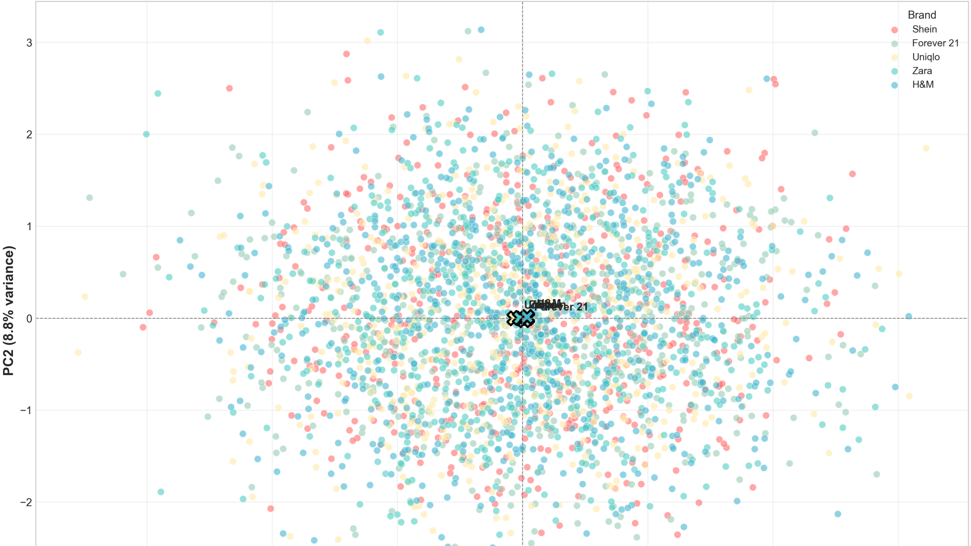


Interpretation:

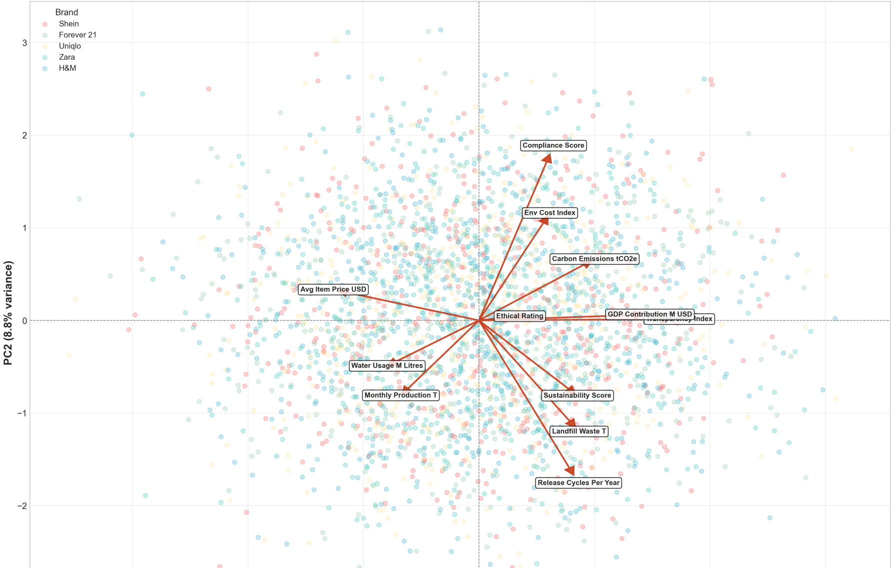
- Arrows → Variable loadings
- Close to circle → Well represented
- Same direction → Positive correlation
- Opposite → Negative correlation

Note: Variables spread across quadrants ⇒ Independence

Individuals Plot: Brands in Principal Component Space
(X markers = Brand Centroids)



Biplot: Variables and Individuals in PC Space



PC1 (9.20% variance)

Environmental Impact Dimension

- High loadings: Carbon emissions, Water usage, Production volume
- Interpretation: Overall environmental footprint

PC2 (8.83% variance)

Sustainability Governance Dimension

- High loadings: Transparency index, Compliance score
- Interpretation: Corporate governance quality

Total variance explained by PC1-PC2: 18.03%

Observation: Eigenvalues are nearly uniform ($\lambda_k \approx 1$)

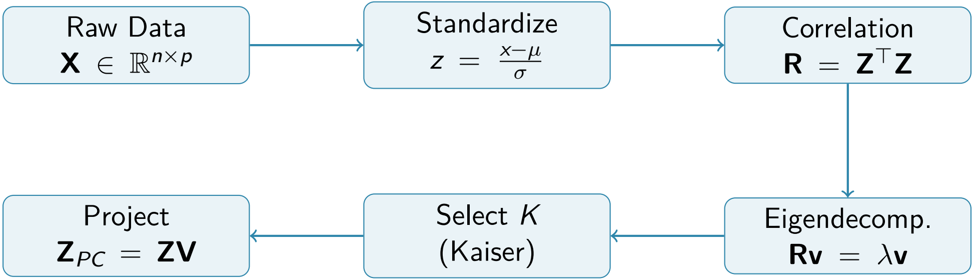
Interpretation

- Variables are **nearly orthogonal** (uncorrelated)
- Each variable captures a **distinct dimension** of sustainability
- **No redundancy** in the measurement system

This is actually informative!

- Environmental impact is **multidimensional**
- Cannot be reduced to a single score
- Each metric provides **unique information**

1. **Dimensionality Reduction:**
   - 12 variables $\rightarrow$ 6 components (Kaiser)
   - 52.51% variance retained
2. **Variable Independence:**
   - Low correlations $\Rightarrow$ metrics capture distinct aspects
   - No major redundancy in environmental data
3. **Brand Analysis:**
   - No significant differentiation between brands
   - Fast fashion industry has homogeneous environmental profile



Tools: Python (NumPy, Pandas, Scikit-learn, Matplotlib)

- **Non-linear extensions:**
  - Kernel PCA for non-linear relationships
  - t-SNE/UMAP for visualization
- **Confirmatory analysis:**
  - Factor Analysis with rotation
  - Structural Equation Modeling
- **Temporal analysis:**
  - Dynamic PCA for trend detection
  - Time series decomposition

Thank You!

Questions?

ST2DA-I2 | 2025-2026

Proof: Variance = Eigenvalue

For principal component $\mathbf{z}_k = \mathbf{Z}\mathbf{v}_k$:

$$\begin{aligned}
\mathrm{Var}(\mathbf{z}_k) &= \frac{1}{n-1}\mathbf{z}_k^\top \mathbf{z}_k = \frac{1}{n-1}(\mathbf{Z}\mathbf{v}_k)^\top(\mathbf{Z}\mathbf{v}_k) \\
&= \frac{1}{n-1}\mathbf{v}_k^\top \mathbf{Z}^\top \mathbf{Z}\mathbf{v}_k = \mathbf{v}_k^\top \mathbf{R}\mathbf{v}_k \\
&= \mathbf{v}_k^\top(\lambda_k \mathbf{v}_k) = \lambda_k \underbrace{\mathbf{v}_k^\top \mathbf{v}_k}_{=1} = \lambda_k \quad \square
\end{aligned}$$

PC	λ	Var.%	Cum.%	Kaiser
1	1.1041	9.20	9.20	✓
2	1.0596	8.83	18.03	✓
3	1.0549	8.79	26.81	✓
4	1.0455	8.71	35.52	✓
5	1.0335	8.61	44.13	✓
6	1.0058	8.38	52.51	✓
7	0.9833	8.19	60.70	
8	0.9762	8.13	68.84	
9	0.9655	8.04	76.88	
10	0.9465	7.88	84.76	
11	0.9299	7.75	92.51	
12	0.8991	7.49	100.00	

1. Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Phil. Trans. R. Soc. A*, 374(2065).
2. Abdi, H., & Williams, L. J. (2010). Principal component analysis. *WIREs Comp. Stats.*, 2(4), 433-459.
3. Pearson, K. (1901). On lines and planes of closest fit. *Phil. Mag.*, 2(11), 559-572.
4. Hotelling, H. (1933). Analysis of a complex of statistical variables. *J. Ed. Psych.*, 24(6), 417-441.
5. Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Ed. & Psych. Meas.*, 20(1), 141-151.