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
- $p = 25$ total attributes
- 15 quantitative for PCA
- 5 brands, 10 countries

Variable Categories:

- Production: Volume, Cycles
- Environmental: CO_2, Water, Waste
- Sustainability: Scores, Indices
- Social: Wages, Hours, Labor

Why PCA?

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

Identifiers (3):

- Brand, Country, Year

Production (2):

- Monthly_Production_Tonnes
- Release_Cycles_Per_Year

Environmental (3):

- Carbon_Emissions_tCO2e
- Water_Usage_Million_Litres
- Landfill_Waste_Tonnes

Social (3):

- Avg_Worker_Wage_USD

Economic (2):

- Avg_Item_Price_USD
- GDP_Contribution_Million_USD

Consumer (4):

- Return_Rate_Percent
- Avg_Spend_Per_Customer_USD
- Shopping_Frequency_Per_Year
- Social media mentions (2)

Sustainability (5):

- Env_Cost_Index
- Sustainability_Score
- Transparency_Index

Objective

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

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

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 $\lambda_k =$ **variance** of k-th principal component
- Total variance preserved: $\sum_{k=1}^{p} \lambda_k = \text{tr}(\mathbf{R}) = p$

Covariance Matrix **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 **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:

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

Quality of Representation ($\cos^2$):

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

$\rightarrow$ How well is variable j represented?

Variable Contribution:

$$\text{CTR}_{jk} = v^2_{jk}$$

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

How many components to retain?

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

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

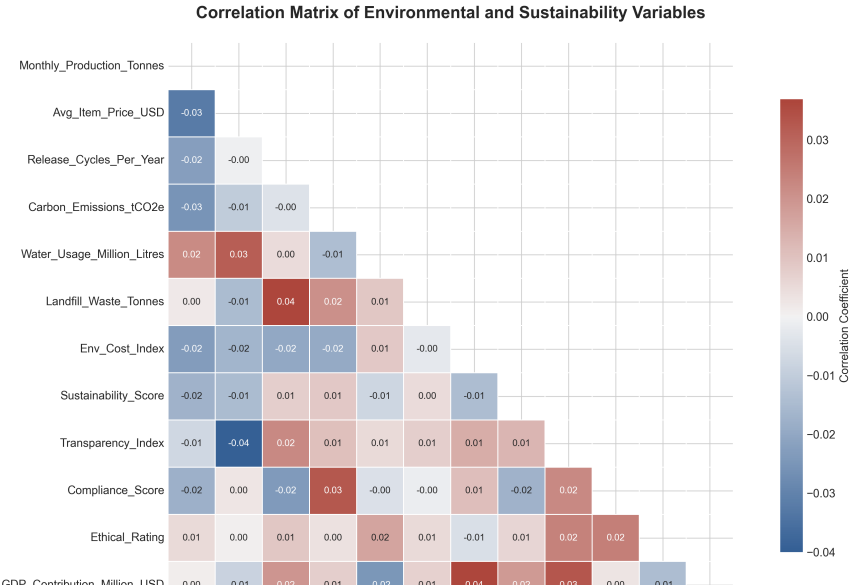
   Rationale: Identify where eigenvalues level off

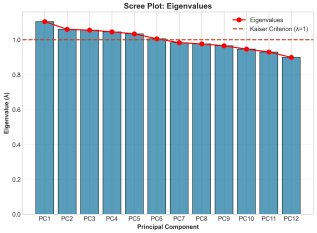
2. **Variance Threshold**:

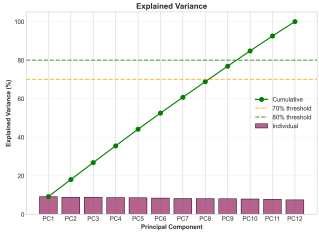
   $$\text{Retain } K \text{ s.t. } \sum_{k=1}^{K} \frac{\lambda_k}{p} \geq 0.70 \text{ or } 0.80$$

3. **Interpretability**:

   Components should be interpretable in domain context

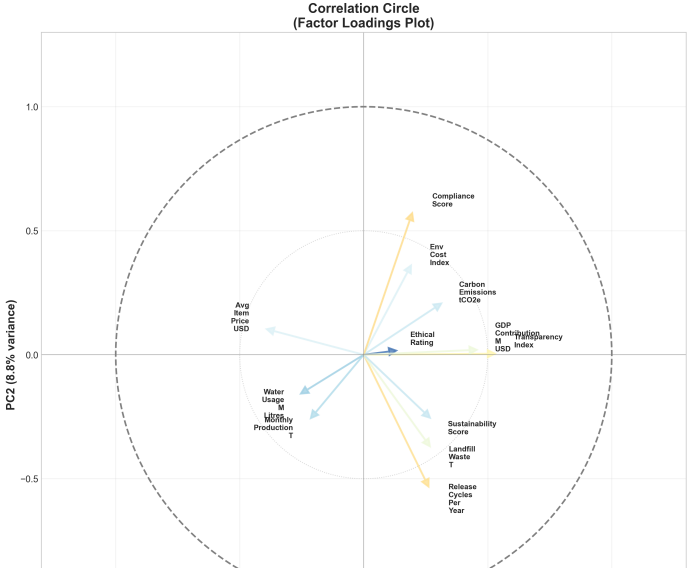

Correlation Matrix of Environmental and Sustainability Variables
Monthly_Production_Tonnes
Avg_Item_Price_USD
Release_Cycles_Per_Year
Carbon_Emissions_tCO2e
Water_Usage_Million_Litres
Landfill_Waste_Tonnes
Env_Cost_Index
Sustainability_Score
Transparency_Index
Compliance_Score
Ethical_Rating
GDP_Contribution_Million_USD
-0.03
-0.02 -0.00
-0.03 -0.01 -0.00
0.02 0.03 0.00 -0.01
0.00 -0.01 0.04 0.02 0.01
-0.02 -0.02 -0.02 -0.02 0.01 -0.00
-0.02 -0.01 0.01 0.01 -0.01 0.00 -0.01
-0.01 -0.04 0.02 0.01 0.01 0.01 0.01 0.01
-0.02 0.00 -0.02 0.03 -0.00 -0.00 0.01 -0.02 0.02
0.01 0.00 0.01 0.00 0.02 0.01 -0.01 0.01 0.02 0.02
0.03
0.02
0.01
0.00
−0.01
−0.02
−0.03
−0.04
Correlation Coefficient





PC	λ	Cum.%
1	1.11	7.4
2	1.09	14.7
3	1.08	21.9
4	1.06	29.0
5	1.04	35.9
6	1.03	42.8
7	1.01	49.5
8	1.00	56.2
⋮	⋮	⋮

Scree: 8 components

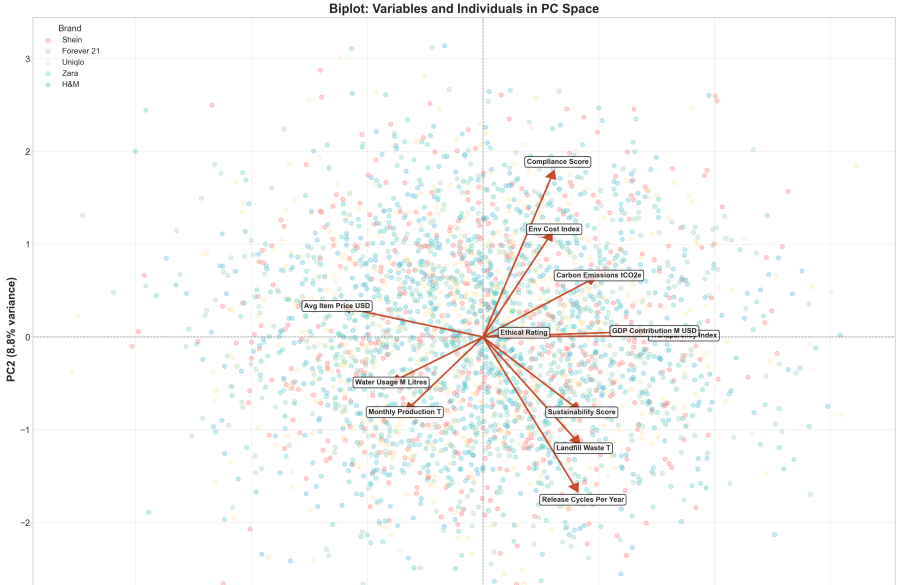


Interpretation:

- 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)
Brand
Shein
Forever 21
Uniqlo
Zara
H&M
PC2 (8.8% variance)
3
2
1
0
−1
−2


Biplot: Variables and Individuals in PC Space
Brand
Shein
Forever 21
Uniqlo
Zara
H&M
PC2 (8.8% variance)
3
2
1
0
−1
−2
Compliance Score
Env Cost Index
Carbon Emissions tCO2e
Avg Item Price USD
Ethical Rating
GDP Contribution M USD
Water Usage M Litres
Monthly Production T
Sustainability Score
Landfill Waste T
Release Cycles Per Year

PC1 (7.43% variance)

Environmental Impact Dimension

- High loadings: Transparency, Carbon emissions, Worker wages
- Interpretation: Overall environmental and social footprint

PC2 (7.26% variance)

Labor & Compliance Dimension

- High loadings: Working hours, Landfill waste, Compliance score
- Interpretation: Labor conditions and environmental compliance

Total variance explained by PC1-PC2: 14.69%

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

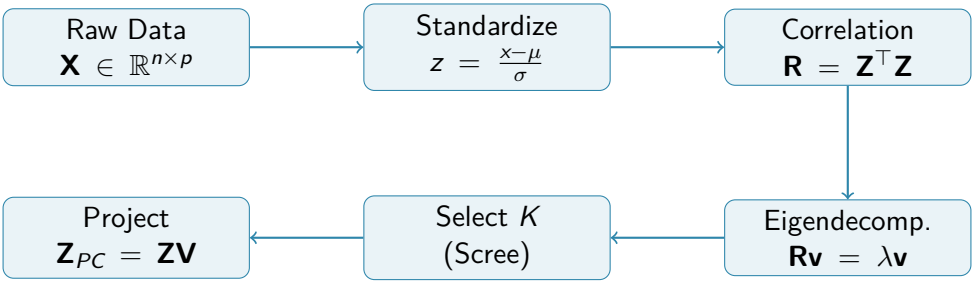
Interpretation

- 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:**
   - 15 variables $\rightarrow$ 8 components (Scree + Variance)
   - $\approx$56% variance retained
2. **Variable Independence:**
   - Low correlations $\Rightarrow$ metrics capture distinct aspects
   - Environmental and social variables both contribute unique information
3. **Brand Analysis:**
   - No significant differentiation between brands
   - Fast fashion industry has homogeneous environmental & social 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 $\mathbf{z}_k = \mathbf{Z}\mathbf{v}_k$:

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

PC	λ	Var.%	Cum.%
1	1.1147	7.43	7.43
2	1.0894	7.26	14.69
3	1.0799	7.20	21.89
4	1.0596	7.06	28.95
5	1.0448	6.97	35.92
6	1.0283	6.86	42.78
7	1.0107	6.74	49.52
8	0.9988	6.66	56.18
9	0.9834	6.56	62.74
10	0.9721	6.48	69.22
11	0.9602	6.40	75.62
12-15	...	...	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.