Report on *A gentle introduction to Principal Component Analysis* (E. Saccenti, 2024)

October 23, 2025

Abstract

This short report summarises the main ideas I found interesting in Edoardo Saccenti's paper "A gentle introduction to principal component analysis using tea-pots, dinosaurs, and pizza", highlights practical applications of PCA the paper demonstrates, and provides reflections on the pedagogical examples used (tea-pots, triceratops, pizza, psoriasis gene expression). The original paper was used as the source for this report. :contentReference[oaicite:0]index=0

1 Overview

Saccenti's paper presents PCA starting from visual intuition (rotations in 2D and 3D) and progressively builds up to the formal linear-algebraic formulation. The paper emphasises PCA as a tool for (a) discovering directions of maximal variance, (b) transforming correlated variables into uncorrelated principal components (scores and loadings), and (c) reducing dimensionality for visualization or downstream analysis. The treatment balances minimal mathematics (trigonometry-based rotation) with practical examples and MATLAB code pointers.

2 Key concepts I found interesting

- **Variance as information.** The paper frames dispersion (variance) along a direction as the information content in that direction. This intuitive link (illustrated with simple point clouds) helps students see why PCA seeks directions that maximize variance.
- **Rotation viewpoint.** Introducing PCA as a sequence of rotations in the coordinate system (plane → space → high dimension) is pedagogically strong: rotations preserve distances but change the variance distribution across axes, making "important" directions explicit.
- **Principal components = linear combinations.** The clear transition from trigonometric rotation formulas in 2D to the eigenvalue/eigenvector solution of the covariance matrix (Appendix A) gives students both intuition and the formal recipe.
- **Loadings and scores distinction.** The exposition distinguishes *loadings* (weights of original variables in each PC) from *scores* (observations projected into PC space), which is crucial for interpretation.

- **When PCA fails / limitations.** The psoriasis gene-expression example demonstrates that PCA is not a universal solution: when variables are mostly uncorrelated (or relationships are nonlinear), PCA may not reveal useful low-dimensional structure.
- **Scaling sensitivity and preprocessing.** The discussion of centering and standardization (unit variance scaling) and their impact on PCA results is a practical, often overlooked, but essential lesson.

3 Applications of PCA illustrated in the paper

The paper uses varied concrete examples to show different PCA uses and limitations:

1. **Shape discovery and visualization (teapot, triceratops):** 2D and 3D coordinate clouds show that PCA (rotation and projection) can reveal an object's intrinsic structure and produce compact visual summaries of complex geometry.
2. **Dimensionality reduction for high-dimensional data:** the synthetic 36,876×1000 example illustrates that a small number of PCs can capture a large fraction of total variance (enabling visualization and storage/compute savings).
3. **Exploratory analysis and clustering (pizza chemistry):** PCA on 300 pizza samples × 7 chemical attributes demonstrated how scores separate brands into groups and how loadings indicate which variables drive differences (e.g., fat, sodium, calories). This is a typical chemometrics/food science use case.
4. **Diagnostics and model limitations (psoriasis gene-expression):** PCA used as a first-step exploratory tool can quickly indicate whether obvious group separation exists; lack of separation can signal either no signal in the measured variables or the need for different methods (nonlinear methods, feature selection, supervised learning).
5. **Multicollinearity and denoising:** by transforming correlated variables into uncorrelated PCs and retaining only the top components, PCA reduces redundancy and can act as a denoising step before regression or classification.

4 Reflections on the specific examples

4.1 Tea-pots

The teapot example is excellent for an introductory audience: it is concrete, visually memorable, and immediately demonstrates the invariance of shape under rotation plus the difference between "seeing" structure in 2D versus higher dimensions. As a pedagogical device, it motivates the need for a method (PCA) that finds informative directions automatically.

4.2 Triceratops (3D → synthetic expansion)

The triceratops example shows two pedagogical points at once: (1) how rotation in 3D can reveal a recognizable structure that was not obvious initially; (2) how expanding a structured low-dimensional object into a higher-dimensional dataset (adding many uncorrelated noise variables) lets PCA compress back to the informative axes. This is a very effective demonstration of dimensionality reduction and denoising.

4.3 Pizza dataset

This is a practical, domain-relevant example (food chemistry) that highlights interpretation: score plots reveal clusters (brands) and loading plots indicate which chemical attributes separate those clusters. The example nicely demonstrates how domain knowledge (e.g., what fat or moisture implies) is needed to interpret PCs meaningfully.

4.4 Psoriasis gene-expression

Arguably the most valuable cautionary example: when variables lack strong linear correlations, PCA will not produce a few dominant PCs and the data cloud appears spherical in PC space. This shows students that PCA is not a magic bullet and underscores the need to inspect loadings/variance-explained and to consider alternative methods (feature engineering, nonlinear dimension reduction, or supervised approaches).

5 Pedagogical strengths and suggested enhancements

Strengths

- Builds intuition first (rotation and variance), then gives the eigen-decomposition formula—good cognitive sequencing.
- Uses memorable, cross-disciplinary examples that appeal to non-statisticians.
- Includes practical pointers (MATLAB commands) and publicly available data/code repositories.

Possible enhancements

- Add a short hands-on R/Python notebook (the paper gives MATLAB) so students who do not use MATLAB can reproduce examples quickly.
- A short interactive widget (2D rotation slider) would make the rotation-to-max-variance idea even more tactile for students.
- Brief demonstration of a nonlinear alternative (e.g., t-SNE or UMAP) on the psoriasis set would reinforce when to prefer nonlinear methods.

6 Technical notes (concise)

- **Centering and scaling:** Centering (subtracting column means) is mandatory for standard PCA; scaling to unit variance is recommended when variables have different units or magnitudes.
- **Computation:** PCA can be performed either via eigendecomposition of the covariance matrix or via singular value decomposition (SVD) of the centered data matrix; SVD is numerically robust for large matrices.
- **Choosing number of components:** Use scree plots, cumulative explained variance thresholds, cross-validation or statistical tests; this is a non-trivial decision and context dependent.

- **Limitations:** PCA captures linear structure (variance along linear directions). It is insensitive to nonlinear manifold structure and to patterns that do not produce large linear variance.

7 Concluding remarks

Saccenti's paper provides an accessible, well-structured introduction to PCA that is particularly well-suited to students with limited linear algebra background. The sequence from visual intuition (teapots, dinosaurs) to formalism (eigenanalysis) is pedagogically effective, and the pizza/psoriasis examples nicely cover both the power and the limits of PCA. For teaching, coupling the paper with interactive visual tools and notebooks in multiple languages (MATLAB, R, Python) would broaden accessibility.

References

References

[1] E. Saccenti, "A gentle introduction to principal component analysis using tea-pots, dinosaurs, and pizza," *Teaching Statistics*, 46:38–52, 2024. (source document used for this report). :contentReference[oaicite:1]index=1