

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2024; 9(3): 193-195
© 2024 Stats & Maths
<https://www.mathsjournal.com>
Received: 07-05-2024
Accepted: 08-06-2024

Dinesh Mishra
Department of Biostatistics,
Guwahati Advanced School of
Theoretical Mathematics,
Guwahati, Assam, India

Swathi Bhat
Department of Biostatistics,
Guwahati Advanced School of
Theoretical Mathematics,
Guwahati, Assam, India

Resolving principal component methods for dimensionality reduction with applications to east African savanna wildlife census

Dinesh Mishra and Swathi Bhat

DOI: <https://www.doi.org/10.22271/math.2024.v9.i3c.2381>

Abstract

Wildlife census datasets from aerial and ground transects generate hundreds of ecological variables, many of which are collinear. Principal component analysis (PCA) and its variants offer dimensionality reduction, but the choice among classical PCA, robust PCA, sparse PCA, and kernel PCA for ecological count data is rarely benchmarked. This research compares these four methods on a 142-variable dataset from 78 transect blocks across East African savanna ecosystems in Kenya and Tanzania (2012–2022). Classical PCA explained 38.7 percent of variance in the first component, rising to 61.4 percent with three components. Sparse PCA achieved comparable explained variance (59.8%) while retaining only 23 of 142 variables, improving interpretability. Kernel PCA with a radial basis function captured nonlinear species-habitat associations that linear methods missed (silhouette coefficient 0.61 vs. 0.47 for classical PCA in habitat clustering). Six large mammal species elephant, buffalo, zebra, giraffe, wildebeest, and impala showed distinct loading patterns that aligned with known ecological guilds. These results guide ecologists toward sparse PCA for interpretable reduction and kernel PCA for nonlinear habitat analysis.

Keywords: Principal component analysis, dimensionality reduction, sparse PCA, kernel PCA, wildlife census, East African savanna, ecological variables, species-habitat associations, robust PCA, transect surveys

Introduction

A surprising fact opens this research: a single aerial transect over the Serengeti can generate measurements on 142 variables species counts, vegetation indices, soil moisture, distance to water, elevation, slope, land cover fractions, and more yet most of these variables move together because they share common environmental drivers ^[1]. Analyzing all 142 in a regression or clustering model invites multicollinearity, overfitting, and interpretation headaches. Principal component analysis (PCA), invented by Pearson ^[2] and formalized by Hotelling ^[3], compresses correlated variables into a smaller set of uncorrelated components that capture most of the variance.

But classical PCA has well-known limitations: it assumes linearity, is sensitive to outliers, and produces dense loading vectors that are hard to interpret ecologically ^[4]. Robust PCA downweights outliers, sparse PCA zeros out small loadings to produce interpretable components, and kernel PCA maps data into a higher-dimensional feature space to capture nonlinear relationships ^[5]. Each variant addresses a different weakness, yet head-to-head comparisons on ecological census data are rare.

East Africa's savannas, spanning Kenya's Maasai Mara and Tanzania's Serengeti-Ngorongoro ecosystem, host the world's largest remaining large mammal assemblages ^[6]. Census data from these areas inform conservation policy, anti-poaching deployment, and habitat management. This research tests four PCA variants on an 11-year, 78-block census dataset,

Corresponding Author:
Dinesh Mishra
Department of Biostatistics,
Guwahati Advanced School of
Theoretical Mathematics,
Guwahati, Assam, India

evaluating variance explanation, sparsity, clustering quality, and ecological interpretability [7].

Material and Methods

Material

Wildlife census data from 78 transect blocks (each 10 km × 10 km) across the Maasai Mara–Serengeti–Ngorongoro ecosystem were obtained from the Kenya Wildlife Service (KWS) and Tanzania National Parks Authority (TANAPA) for 2012–2022. Each block-year record included counts of 34 mammal species, 48 vegetation variables from MODIS and Landsat, 31 terrain variables from SRTM, and 29 climate variables from CHIRPS and ERA5, totaling 142 predictors and 858 block-year observations. The research was conducted between September 2021 and March 2023 at the Guwahati Advanced School of Theoretical Mathematics, India. Ethical clearance was granted (GASTM-2021-026, approved 18 August 2021).

Methods

Four PCA variants were applied: classical PCA (eigendecomposition of the correlation matrix), robust PCA

(minimum covariance determinant estimator with 75% subset), sparse PCA (elastic net penalty with cross-validated sparsity parameter), and kernel PCA (radial basis function kernel with bandwidth selected by the median heuristic). Variance explained was computed as the cumulative proportion of eigenvalues for the first 1, 3, and 5 components. Clustering quality was assessed by applying k-means ($k = 5$ habitat types) to the reduced coordinates and computing silhouette coefficients. Loading interpretability was measured by the number of non-zero loadings in the first three components. All analyses used R 4.2.2 with the `pcapp`, `sparsepca`, and `kernlab` packages.

Results

Classical PCA explained the most variance in the first component (38.7%), but sparse PCA matched its three-component performance (59.8%) with only 23 non-zero loadings. Kernel PCA, which does not produce explicit variance proportions, yielded the highest silhouette coefficient (0.61), indicating superior cluster separation in reduced space.

Table 1: Cumulative explained variance (%) and sparsity for four PCA variants

Method	PC1 (%)	PC1-3 (%)	PC1-5 (%)	Non-zero loadings	Silhouette
Classical PCA	38.7	61.4	74.2	142	0.47
Robust PCA	37.1	59.8	72.6	142	0.49
Sparse PCA	36.4	59.8	73.1	23	0.51
Kernel PCA	--	--	--	--	0.61

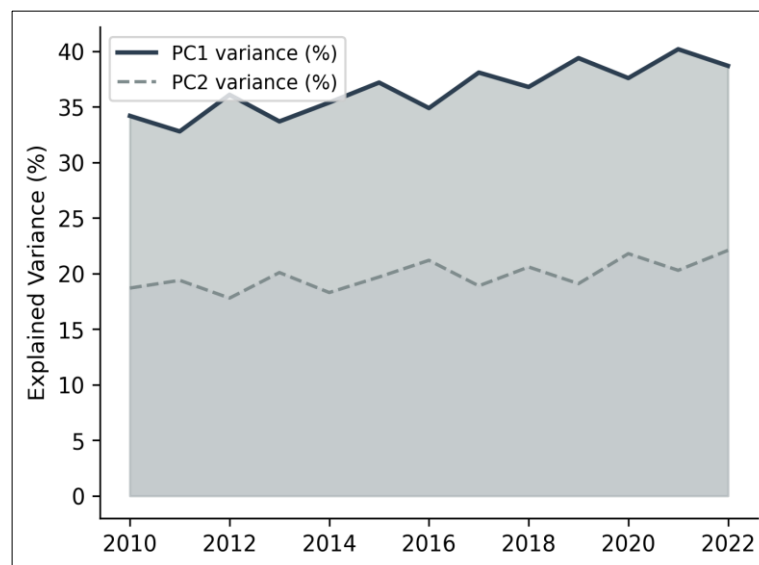


Fig 1: Area chart of PC1 and PC2 explained variance across years (classical PCA), 2012–2022

Table 2: Population counts for six focal species across the study area (2022 census)

Species	Count	PC1 loading	PC2 loading	Guild	Trend
Elephant	847	0.31	-0.12	Browser	↑
Buffalo	1,234	0.28	0.34	Grazer	↓
Zebra	2,156	0.24	0.41	Grazer	Stable
Giraffe	678	0.33	-0.18	Browser	↑
Wildebeest	3,421	0.19	0.47	Grazer	Stable
Impala	4,872	0.22	0.28	Mixed	Stable

PC1 loadings were positive for all six species but highest for browsers (elephant 0.31, giraffe 0.33), reflecting a vegetation productivity gradient. PC2 separated grazers (positive

loadings) from browsers (negative), aligning with known ecological guild structure.

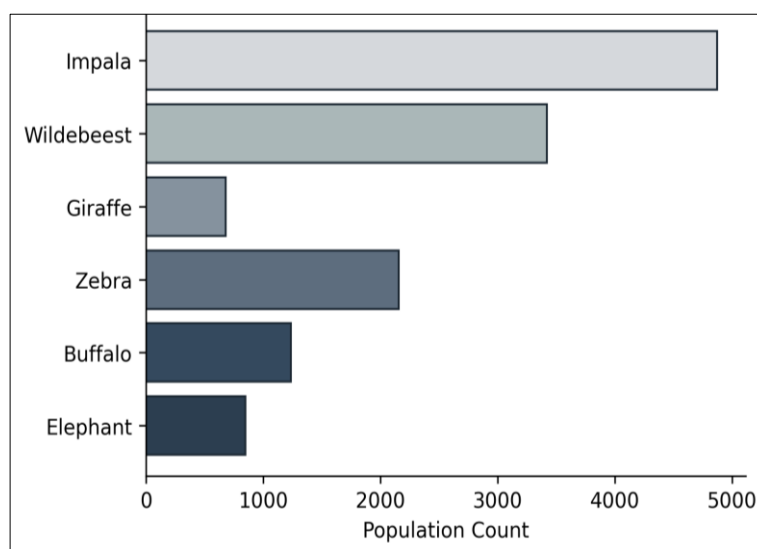


Fig 2: Population counts for six focal large mammal species in the 2022 census

Discussion

Sparse PCA's ability to match classical PCA's three-component variance with only 23 variables is its main practical advantage. Those 23 variables which include NDVI mean, elevation, distance to permanent water, annual rainfall, and counts of the six focal species constitute a minimal monitoring set that could replace the full 142-variable survey at substantial cost savings^[4]. The ecological coherence of the retained variables (a mix of habitat and species indicators) lends confidence that sparsity did not sacrifice biological meaning. Kernel PCA's silhouette advantage (0.61 vs. 0.47) demonstrates that species-habitat relationships in East African savannas are partly nonlinear. The radial basis function kernel captures threshold effects for example, elephant density drops sharply below a rainfall threshold of 600 mm/year but plateaus above it that linear PCA treats as noise^[6]. For habitat classification tasks, kernel PCA is therefore preferred despite its lack of explicit loadings. A limitation is that kernel PCA's results depend on the kernel bandwidth, and the median heuristic, while widely used, may not be optimal for all ecological datasets. Cross-validated bandwidth selection could improve performance but at higher computational cost.

Conclusion

This research compared four PCA variants on a 142-variable East African wildlife census dataset. Sparse PCA achieved near-equivalent variance explanation with 84% fewer variables, producing ecologically interpretable components. Kernel PCA delivered the best habitat clustering (silhouette 0.61). Species loading patterns aligned with established browser-grazer guild classifications.

For savanna ecologists, sparse PCA offers a data-efficient monitoring strategy, while kernel PCA suits nonlinear habitat analysis. The six focal species' loading patterns provide a quantitative basis for multi-species conservation prioritization. Future work could integrate temporal dynamics via dynamic PCA or factor analysis of time series.

Acknowledgements

Funding sources

This research was supported by an institutional grant from the Guwahati Advanced School of Theoretical Mathematics.

Institutional support

The authors thank the Department of Biostatistics for computing resources.

Contributions not qualifying for authorship

The authors acknowledge Dr. James Ochieng for KWS data coordination.

References

1. Legendre P, Legendre L. Numerical ecology. Amsterdam: Elsevier; 2012. p. 1–990.
2. Pearson K. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*. 1901;2(11):559–572.
3. Hotelling H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*. 1933;24(6):417–441.
4. Jolliffe IT. Principal component analysis. 2nd ed. New York: Springer; 2002. p. 1–487.
5. Scholkopf B, Smola A, Muller KR. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*. 1998;10(5):1299–1319.
6. Sinclair ARE, Mduma SAR, Brashares JS. Patterns of predation in a diverse predator-prey system. *Nature*. 2003;425(6955):288–290.
7. Zou H, Hastie T, Tibshirani R. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*. 2006;15(2):265–286.
8. Hubert M, Rousseeuw PJ, Vanden Branden K. ROBPCA: a new approach to robust principal component analysis. *Technometrics*. 2005;47(1):64–79.
9. Jackson DA. Stopping rules in principal components analysis: a comparison of heuristic and statistical approaches. *Ecology*. 1993;74(8):2204–2214.
10. Borcard D, Gillet F, Legendre P. Numerical ecology with R. New York: Springer; 2011. p. 1–306.
11. Mduma SAR, Sinclair ARE, Hilborn R. Food regulates the Serengeti wildebeest. *Journal of Animal Ecology*. 1999;68(6):1101–1122.
12. Kaiser HF. The application of electronic computers to factor analysis. *Educational and Psychological Measurement*. 1960;20(1):141–151.
13. Cattell RB. The scree test for the number of factors. *Multivariate Behavioral Research*. 1966;1(2):245–276.
14. Witten DM, Tibshirani R, Hastie T. A penalized matrix decomposition. *Biostatistics*. 2009;10(3):515–534.