ESP IJAST
ESP International Journal of Advancements in Science & Technology
ISSN: 2583-9233 / Volume 2 Issue 3 July 2024 / Page No: 20–28
Paper Id: IJAST-V2I3P103 / Doi: 10.56472/25839233/IJAST-V2I3P103

Original Article

Developments in Artificial Intelligence and Machine Learning: Recent Advances and Prospect

Gothatamang Patrick Nthoiwa[1], Ramasaymy Sivasamy[2]

[1,2]*Department of Statistics, University of Botswana, Gaborone Botswana.*

Received Date: 23 May 2024 Revised Date: 20 June 2024 Accepted Date: 22 July 2024

Abstract: *Principal Component Analysis (PCA) is a powerful tool for understanding the underlying structure and relationships within multivariate datasets, often collected through extensive field surveys and monitoring programs. This study explores the best practices for performing PCA on incomplete datasets with missing values, with a focus on the significance of sophisticated imputation techniques or resilient missing data strategies to maintain the analytical value of ecological datasets. The study proposes leveraging the power of Singular Value Decomposition (SVD) and the inherent low-rank structure of ecological data, offering a robust framework for analyzing complex ecological systems, enabling the identification of latent ecological factors, prediction of missing observations, and ultimately, a deeper understanding of the dynamics governing these systems. The PCA results conducted on a simulated dataset illustrate the performance comparison between two different methods for handling missing data in PCA. The NIPALS method, while offering an alternative standardization approach, should be used with caution due to its potential to significantly alter the PCA outcomes. Regularized SVD demonstrated the most consistent performance across all levels of missingness, indicating its robustness for handling the missing data. Future research should explore alternative etiologies and their effects on PCA outcome, as well as sensitivity analyses to determine optimal regularization parameters.*

Keywords: *Incomplete Datasets, Low-Rank, Matrix Completion, Principal Component Analysis (PCA), Regularization, Singular Value Decomposition (SVD).*

I. INTRODUCTION

Analyzing ecological data is crucial to understanding the complex interactions within natural systems. As the volume and complexity of environmental data continue to grow, innovative data processing and analysis techniques have become increasingly important. One promising approach is matrix completion using regularized principal component analysis (PCA) (Shen & Huang, 2008) to address the challenge of missing data in ecological datasets [1].

Ordination in ecology is a fundamental statistical approach to understanding the relationships between species, sites, and environmental variables [2]. Principal Component Analysis (PCA) offers a robust statistical framework to achieve this by reducing the dimensionality of the data, identifying the principal axes of variation, and visualizing patterns [3]. PCA simplifies multivariate relationships by transforming the original variables into principal components that capture the most variation in the data. By analyzing these principal components, ecologists can interpret the ecological processes and factors driving the observed patterns.

The role of PCA in ecological analysis is multifaceted [3]. First, PCA reduces the complexity of ecological data by transforming it into a set of principal components, which are orthogonal vectors representing the directions of maximum variance [3]. This reduction helps focus on the most significant patterns and relationships, crucial for interpreting ecological data and identifying key factors influencing species distributions and community structures [3]. Second, the principal components identified by PCA reveal the significant axes of variation within the data, often corresponding to underlying ecological gradients or patterns, such as variations in species composition along environmental gradients like moisture, temperature, or elevation. Understanding these axes allows ecologists to infer the ecological processes and factors that shape the observed patterns [4, 5].

Third, PCA facilitates the creation of ordination plots, graphical representations of the relationships among samples and variables. In these plots, samples, different sites) are represented as points, and their positions reflect their similarities or differences, enabling the visualization of eco-logical patterns. Missing values in ecological data matrices can introduce significant bias and uncertainty, hindering the accuracy of pnt analysis (PCA) and the subsequent accuracy of modeling ecological patterns and processes [6]. Missing data is a ubiquitous issue in empirical research, with far-reaching implications for the validity and reliability of study findings. Rubin's work has identified three fundamental types of missing data mechanisms: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR) [7].

[8] MCAR, the most benign form, occurs when the probability of missing data is unrelated to observed and unobserved variables. Conversely, MAR depends on observed data but not the unobserved missing values, posing a more challenging scenario. MNAR, the most problematic case, arises when missingness depends on unobserved missing values, even after accounting for observed data [9]. Principled approaches such as multiple imputation and complete information maximum likelihood have been proposed to address this challenge [10] [9].

Multiple imputations, in particular, have been shown to produce efficient and valid estimates when the missing data mechanism is correctly specified. However, applying these techniques is not yet widespread in the ecological research community.[11][12–15]. Little and Rubin's work on missing data identifies three fundamental types [7]: Missing Completely at Random (MCAR), missing at Random (MAR), and Missing Not at Random (MNAR). MCAR is the most benign type, where the probability of missing data is unrelated to observed and unobserved values. Conversely, MAR has MARbserved data and is more challenging as it depends on unobserved missing values. For example, if bird count data is missing due to logistical difficulties but is recorded in the data, MAR allows valid inferences under certain conditions. MNAR, on the other hand, depends on unobserved missing values, even after accounting for observed data.

This type of missing data is the most challenging, as standard methods may lead to biased results. The imputation method estimates individual missing values within a dataset, often relying on statistical models or relationships between observed variables. Techniques include mean imputation, regression imputation, multiple imputation, and k-nearest neighbors imputation. The choice of imputation method depends on the data type, missing data mechanism, and desired properties of the imputed values [14, 16].

This method assumes that the ecological data is a multivariate normal distribution, that each variable is normally distributed, and that the relationship between variables is also normal. Parametric methods assume a specific distribution for the data and estimate the parameters of this distribution to impute missing values. The choice between these approaches depends on data characteristics, the number of missing values, computational complexity, and the complexity of the data.[17] We can take advantage of the low-rank structure that many ecological systems have by expressing ecological data. This low-rank property suggests that a small number of underlying variables, such as species interactions or environmental gradients, are responsible for the observed patterns [18, 19]. Then

$$Y \approx L \times F$$

Where:

- Y (n x p): The observed data matrix (e.g., grass species).
- L (n x r): A matrix representing the site-specific effects of the underlying ecological factors.
- F (r x p): A matrix representing the year-specific effects of the underlying ecological factors.
- r: The rank of the approximation, representing the number of underlying ecological factors. This is typically much smaller than both n and p.

Appropriate modifications may be made to current algorithms when using PCA techniques to handle missing data and increase the accuracy of the results. For example, using regularized singular value decomposition (SVD) in PCA eliminates some of the drawbacks of conventional PCA and produces more accurate, robust, and consistent findings.

The best practices for performing PCA on incomplete datasets are suggested here, focusing on the significance of resilient missing data strategies to maintain the analytical value of ecological datasets. These techniques involve generalized low-rank models (GLRMs), which employ low-rank matrices to estimate any given data array.[20]

The structure of ecological data with missing values is analogous to the problem encountered in recommender systems, where missing entries in user-item matrices must be predicted.[21, 22] In recommender systems, Cand`es' groundbreaking work on matrix completion provides a robust framework for addressing this challenge[5, 23, 24]. Matrix completion, a technique rooted in mathematics and machine learning, offers a promising solution. By leveraging the inherent structure and correlations present within ecological data, matrix completion aims to estimate the missing values in a principled manner.

This approach is particularly well-suited for low-rank datasets, where a few latent factors can represent the underlying data. In the ecological context, these latent factors could correspond to underlying environmental gradients, species interactions, or other driving forces shaping the observed patterns. Matrix completion techniques have been developed to effectively handle missing data by leveraging the low-rank nature of the dataset.

Table 1: Comparison of Low-Rank Assumption and Multivariate Normal Distribution Approaches

Feature	Low-Rank Assumption	Multivariate Normal Distribution
Underlying Assumption	Data has a low-rank structure.	Data follows a multivariate normal distribution.
Missing Data Imputation	Matrix Completion (e.g., nuclear norm minimization, ALS)	Parametric Methods (e.g., EM, regression imputation)
Strengths	Works well for high-dimensional data with strong correlations. Robust to outliers.	Interpretable and straightforward if the normality assumption holds.
Limitations	It may not be suitable for data with complex, non-linear relationships.	Sensitive to outliers and deviations from normality

The study aims to improve the accuracy and effectiveness of Principal Component Analysis (PCA) in scenarios with incomplete data. Traditional PCA, often implemented through Singular Value Decomposition (SVD), is effective for dimensionality reduction and feature extraction but can be significantly hampered by missing values. The researchers propose an alternative approach that uses mathematical relationships between matrix decomposition techniques to develop a modified PCA algorithm that can handle missing data while maintaining PCA's core principles. This innovative approach could significantly impact fields such as bioinformatics, social sciences, and engineering, where incomplete datasets are standard.

II. DATA AND METHODS

Assessing PCA Performance under Various Missing Data Scenarios:

A low-rank ecological dataset was simulated and used as the ground truth to assess PCA performance under various missing data scenarios. PCA was initially applied to this complete dataset, calculating singular values and variances for later comparison.

Missing data was introduced at 10%, 30%, and 70%. This was done randomly across the dataset.
Three distinct PCA approaches were compared on these incomplete datasets:

A. Statistical methods:

Standard Principal Component Analysis (PCA) using Singular Value Decomposition (SVD) involves preprocessing the data, extracting the right singular vectors, and calculating the proportion of variance explained by each PC.

Principal Component Analysis (PCA) using Singular Value Decomposition (SVD):

a) Given:

- A data matrix **X** of size $n \times p$, where:
- n is the number of observations (samples)
- p is the number of variables (features)

i) Singular Value Decomposition (SVD): Decompose the centered data matrix $\tilde{X}$:
using SVD:

$$\tilde{X} = U\Sigma V^T$$

Where:

- **U** is an $n \times n$ orthogonal matrix (left singular vectors)
- **Σ** is an $n \times p$ diagonal matrix of singular values ($\sigma_1 \geq \sigma_2 \geq . . . \geq \sigma_p \geq 0$)
- **V** is a $p \times p$ orthogonal matrix (right singular vectors)

ii) Principal Components (PCs):

- The columns of **V** are the data's principal components (PCs).
- The first PC (the first column of **V**) represents the direction of maximum variance in the data.
- The second PC represents the direction of the second largest variance, orthogonal to the first PC, and so on.

iii) Scores:

- The scores are the projections of the original data points onto the principal components. They are calculated as follows:

$$Scores = \boldsymbol{\tilde{X} V}$$

- Alternatively, they can be calculated as:

$$Scores = \boldsymbol{U\Sigma}$$

iv) Explained Variance:

- The singular values in **Σ** are related to the variance explained by each principal component.
- The proportion of total variance explained by the i-th PC is:

Dimensionality Reduction:

$$\frac{\lambda i}{\Sigma\ \lambda j}$$

Where $\lambda i = \sigma^2{}_i$

To reduce the dimensionality of the data, select the top k principal components (columns of **V**), where $k < p$. This is equivalent to keeping the top k singular values and vectors in the SVD.

- The reduced data matrix can be reconstructeollows:

$$\boldsymbol{X}^{\sim}{}_{\boldsymbol{k}} = \boldsymbol{U}_{\boldsymbol{k}}\boldsymbol{\Sigma}_{\boldsymbol{k}}\boldsymbol{V}_{\boldsymbol{k}}{}^{\boldsymbol{T}}$$

Where $\mathbf{U}_k$, $\mathbf{\Sigma}_k$, and $\mathbf{V}_k$ are truncated versions of **U**, **Σ**, and **V** respectively.

The Eckart-Young theorem guarantees that the low-rank approximation obtained by truncating the SVD is the best possible approximation for minimizing the squared reconstruction error. NIPALS for PCA is used to iteratively compute principal components to maximize captured variance in the presence of missing data.

B. NIPALS Algorithm (R):

PCA was performed directly on the data with missing values, utilizing the NIPALS algorithm available in R. This approach does not require prior imputation.[25]

C. Imputation Methods:

Two common imputation techniques were applied before conducting standard PCA:

- **Mean Imputation:** Missing values were replaced with the mean of their respective variables.
- **Multiple Imputation:** Several plausible imputed datasets were generated, and PCA was applied to each, with results pooled for analysis.[26]

D. Regularized PCA:

A regularized PCA method was employed to handle missing data directly without requiring imputation.

For each method and missingness level, the resulting principal components (PCs) and their associated variances were compared to the results obtained from the complete dataset. This comparison allowed for assessing how well each method preserved the original data structure and relationships in the presence of varying amounts of missing data.

III. REGULARIZED SVD FOR PCA TO HANDLE MISSING DATA

A. Given:

- A data matrix **X** of size $n \times p$ with missing values.

B. Steps:

a) Centering:

- Subtract the mean of each column (feature) from the corresponding column in **X**, ignoring the missing values.
- This yields a centered data matrix.

$$\boldsymbol{X}^{\sim} = \boldsymbol{X} - mean(\boldsymbol{X})$$

b) Initialize Missing Values:

- Initialize missing values in $X^{\sim}$ estimates with zeros, column means, or some other *initial*

c) Regularized SVD:

- Decompose the centered data matrix using regularized SVD:

$$\boldsymbol{X}^{\sim} = \boldsymbol{U\Sigma V^T}$$

- Regularization is applied to prevent overfitting and handle the missing data. This is achieved by adding a regularization term to the SVD optimization problem:

$$_{\boldsymbol{U,\Sigma,V}}min\ ||\boldsymbol{W} \odot (\boldsymbol{X}^{\sim} - \boldsymbol{U\Sigma V}T\ )||^2{}_F + \lambda(\ ||\boldsymbol{U}||_{\boldsymbol{F}}{}^{\mathbf{2}} + |\boldsymbol{V}||^{\mathbf{2}}\ )_F$$

Where:

- **W** is a weight matrix with w_{ij} = 0 if x_{ij} is missing and w_{ij} = 1 otherwise.
- $\odot$ denotes element-wise multiplication.
- $\|\cdot\|_F$ denotes the Frobenius norm.
- λ is the regularization parameter.

d) Iterative Imputation:

- Use the resulting U, Σ, and V to reconstruct *and* update missing values
- Repeat the Regularized SVD and imputation steps until convergence or a maximum number of iterations is reached.

e) Principal Components (PCs):

- The columns of **V** are the data's principal components (PCs).
- The first PC (the first column of **V**) represents the direction of maximum variance in the data.
- The second PC represents the direction of the second largest variance, orthogonal to the first PC, and so on.

f) Scores:

- The scores are the projections of the original data points onto the principal components. They are calculated as follows:

$$\text{Scores} = \tilde{\mathbf{X}}\,\mathbf{V}$$

- Alternatively, they can be calculated as:

$$\text{Scores} = \mathbf{U\Sigma}$$

g) Explained Variance:

- The singular values in **Σ** are related to the variance explained by each principal component.
- The proportion of total variance explained by the i-th PC is:

$$\frac{\lambda i}{\Sigma\ \lambda j}$$

$$Where\ \lambda i = \sigma^2{}_i$$

C. Dimensionality Reduction:

- To reduce the dimensionality of the data, select the top k principal components (columns of **V**), where $k < p$. This is equivalent to keeping the top k singular values and vectors in the SVD.
- The reduced data matrix can be reconstructed as follows:

$$\tilde{\boldsymbol{X}}_{\boldsymbol{k}} = \boldsymbol{U}_{\boldsymbol{k}}\boldsymbol{\Sigma}_{\boldsymbol{k}}\boldsymbol{V}_{\boldsymbol{k}}{}^{\boldsymbol{T}}$$

$\mathbf{U}_k$, $\mathbf{\Sigma}_k$, and $\mathbf{V}_k$ are truncated versions of **U**, **Σ**, and **V** respectively.
The performance of each method will be evaluated by comparing explained variance, reconstruction error,

IV. RESULTS

A. Simulated Data:

Assessing PCA Performance under Various Missing Data Scenarios:

To assess PCA performance under various missing data scenarios, a low-rank ecological dataset was simulated and used as the ground truth. PCA was initially applied to this complete dataset, calculating singular values and variances for later comparison.

Missing data was introduced at 10%, 30%, and 70%. This was done randomly across the dataset.
Three distinct PCA approaches were compared on these incomplete datasets:

B. NIPALS Algorithm (R):

PCA was performed directly on the data with missing values, utilizing the NIPALS algorithm available in R. This approach does not require prior imputation.

C. Imputation Methods:

Two common imputation techniques were applied before conducting standard PCA:

- Mean Imputation: Missing values were replaced with the mean of their respective variables.
- Multiple Imputations: Several plausible imputed datasets were generated, and PCA was applied to each, with results pooled for analysis.

D. Regularized PCA:

A regularized PCA method was employed to handle missing data directly without requiring imputation. For each method and missingness level, the resulting principal components (PCs) and their associated variances were compared to the results obtained from the complete dataset. This comparison allowed for assessing how well each method preserved the original data structure and relationships in the presence of varying amounts of missing data. The table below summarizes the comparison metrics for each technique and missingness level.

Table 2 presents a comparative analysis of three methods for handling missing data in Principal Component Analysis (PCA): NIPALS, Mean Imputation, and Regularized SVD. The evaluation focuses on preserving the original data structure, as measured by differences in explained variance (for the first two principal components, PC1 and PC2) and correlations between loadings across varying levels of missing data (10%, 30%, and 70%).

E. Nipals:

NIPALS demonstrates an increasing deviation from the original PCA results as the percentage of missing data increases. While loadings correlations remain relatively high (0.89 or above), the differences in explained variance become more pronounced, especially at 70% missingness (0.31 for PC1, 0.08 for PC2). This suggests that while NIPALS captures some of the original variable relationships, it struggles to accurately represent the overall variance structure when faced with substantial missing data.

Table 2: Comparison Metrics for Each Method and Missingness Level

Method	Missingness	Explained Variance Diff (PC1, PC2)	Loadings Diff (Correlation)
NIPALS	10%	(0.02, 0.03)	0.98
	30%	(0.24, 0.02)	0.95
	70%	(0.31, 0.08)	0.89
Imputation (Mean)	10%	(0.005, 0.01)	0.99
	30%	(0.01, 0.02)	0.98
	70%	(0.05, 0.06)	0.92
Regularized SVD	10%	(0.001, 0.003)	0.999
	30%	(0.008, 0.015)	0.995
	70%	(0.02, 0.04)	0.97

F. Mean Imputation:

Mean imputation performs well at low to moderate levels of missingness (10% and 30%), with minimal differences in explained variance and very high correlations of loadings (0.98 or above). However, its performance deteriorates at 70% missingness, where the explained variance differences increase (0.05 for PC1, 0.06 for PC2), and the loadings correlation drops to 0.92. This indicates that mean imputation may introduce bias and distort the data structure when many values are missing.

G. Regularized SVD:

Regularized SVD consistently outperforms both NIPALS and mean imputation across all levels of missingness. It exhibits the smallest differences in explained variance and maintains the highest correlations of loadings, even at 70% missingness (0.02 for PC1, 0.04 for PC2; correlation of 0.97). This robustness suggests that regularized SVD is the most reliable method for handling missing data in PCA, as it effectively preserves the original data structure and relationships between variables.

This analysis underscores the importance of carefully selecting a missing data handling method for PCA, as the choice can significantly impact the interpretation and reliability of the results.

H. Real data:

The data is from ecological data on grass species from various zones, focussing on diversity and density. The dataset contains measurements from 62 locations, including species such as Panicum maximum, Cymbopogon caesius, and Eragrostis rigidior. However, significant data gaps exist, including CO3 and HCO3 alkalinity, calcium levels, and trace metal concentrations. The report also emphasizes the variation in species density across locations, indicating different ecological niches. Missing data, such as 58 for Chloris virgata, reduces the reliability of conclusions.

V. MISSING DATA ANALYSIS

A. Missing Data Analysis for Ecological Data:

The following table provides a detailed breakdown of missing data in the ecological dataset. It lists each variable, the statistics or values, the frequency of valid observations, a graphical representation (denoted by bars for simplicity), and the count of valid and missing observations. Below is a summary of the missing data percentages for each variable in the ecological dataset:

Table 3: Percentage of Missing Data by Variable

No	Variable	Missing Percentage (%)
1	ecological zone	0

2	ranch communal	0
3	panicum maximum	61.3
4	cymbopogoncaesius	98.4
5	eragrostisrigidior	43.5
6	eragrostispallens	98.4
7	triraphisschinzii	95.2
8	senyane	98.4
9	anthephorapubescens	98.4
10	eragrostislehmanniana	79.0

The Principal Component Analysis (PCA) was performed on an ecological dataset with missing values using the NIPALS algorithm. The analysis focused on extracting the first five principal components. The results showed that the PCA did not perform as expected due to the high level of missing data and its distribution across variables. The zero eigenvalues suggest that the variance explained by the components is negligible, highlighting the challenges of analyzing incomplete datasets. The first component explains some variance, but subsequent components do not contribute further. Each component required two iterations for computation. Centering and scaling factors were calculated for each variable and standardized before analysis. The scores were undefined for all observations and components, indicating an inability to position the samples in the new component space.

B. Ordinary PCA:

- **PC1** explained 34.6% of the variance, indicating a strong pattern or gradient in the data corresponding to this principal axis.
- **PC2** and **PC3** together accounted for an additional 42.0% of the variance, cumulatively reaching over 76.3% with the first three components.
- The remaining components progressively contributed less, with the total explained variance approaching 100% by the 15th component.

C. Comparison of Explained Variance Differences for PCA Methods Handling Missing Data:

This table compares the explained variance of each principal component (PC) between different PCA methods and the baseline regularized PCA.

Table 4 Explained Variance Differences

Method	**PC1**	**PC2**	**PC3**	**PC4**
Imputed vs. Reg.	0.156	0.149	0.084	0.028
NIPALS vs. Reg.	0.223	0.176	0.042	0.037

This table compares the differences in explained variance for each principal component (PC) between two PCA methods that handle missing data (Imputed and NIPALS) and a baseline regularized PCA (Reg.). The explained variance quantifies how much of the total variation in the data is captured by each PC.

D. Imputed vs. Reg.:

- Positive Differences: The positive values indicate that the PCA performed on the imputed data explains more variance than the regularized PCA for each of the first few PCs (PC1 through PC4).
- Potential Implication: This suggests that the imputation process might have introduced some artificial structure or patterns into the data that were not initially present. These artificial patterns can inflate the variance explained by the PCs.
- Magnitude: The differences are relatively significant for the first two PCs (0.156 and 0.149) and decrease for subsequent PCs. This indicates that the impact of imputation is most pronounced in the primary axes of variation.

E. NIPALS vs. Reg.:

- Positive Differences: Similar to the imputed data, the NIPALS PCA also explains more variance than the regularized PCA for the first few PCs.
- Potential Implication: This implies that the NIPALS standardization procedure, designed to handle missing data, might also introduce some degree of artificial structure or exaggerate existing patterns in the data.
- Magnitude: The differences in explained variance for NIPALS are even more significant than those for the imputed data, especially for PC1 (0.223) and PC2 (0.176). This suggests that NIPALS has a more substantial influence on the variance structure than simple imputation.

F. Imputed vs. Reg:

This row shows the differences in explained variance between PCA performed on imputed data, and PCA performed on the regularized data. The values are positive, indicating that the imputed PCA explained more variance than the regularized PCA for each of the first few PCs. This suggests that imputation might have introduced some artificial structure into the data, leading to slightly inflated variance estimates.

G. NIPALS vs. Reg:

This row compares NIPALS PCA to regularized PCA. The differences are even more significant than those seen with imputation, particularly for PC1 and PC2. This implies that the NIPALS standardization procedure had a more substantial impact on the variance structure of the data than simple imputation.

H. Explained Variance Differences:

Table 5: Explained Variance Differences

Method	PC1	PC2	PC3	PC4
Imputed vs. Reg.	0.156	0.149	0.084	0.028
NIPALS vs. Reg.	0.223	0.176	0.042	0.037

VI. LOADINGS CORRELATION DIFFERENCES

The table presents the correlation differences between the loadings (variable contributions) of each PC for different PCA methods compared to regularized PCA. The results show that for most PCs, the correlation differences are relatively small, indicating that imputed PCA and regularized PCA generally identify similar patterns of variable contributions. However, there are a few exceptions (e.g., PC4) where the difference is more substantial, suggesting that imputation might have altered the relative importance of some variables.

The correlation differences are more significant here compared to the imputed PCA, implying that the NIPALS standardization had a more significant impact on how variables contribute to each PC. The higher differences in the first few PCs suggest that the core structure of the data might be perceived differently under NIPALS compared to regularized PCA.

The results highlight the importance of methodological sensitivity, the sub- the impact of imputation, and the strong influence of NIPALS standardization. Researchers must carefully consider the choice of PCA method and missing data handling technique, as these decisions can substantially affect the interpretation of results. Regularized PCA appears to be relatively robust to missing data, making it a good baseline for comparison. In cases where imputation or NIPALS standardization is used, it’s crucial to assess standardization of the results to these methods and interpret the findings with caution.

The study reveals significant differences in explained variance and loadings correlations between different methods for plant abundance data analysis (PCA). Regularized PCA is a valuable baseline for comparison, as it directly addresses missing values by replacement and centering. Imputed PCA fills in missing values but introduces uncertainty, potentially leading to deviations. The NIPALS standardization method results in the most pronounced differences, possibly due to its different scaling approach and iterative nature. The analysis emphasizes the importance of careful consideration when choosing a method for handling missing values in PCA. Further investigation could explore additional imputation methods, assess the sensitivity of regularized PCA to different parameters, and understand the specific mechanisms behind missing values to determine the most appropriate approach.

Table 6: Loadings Correlation Differences

Method	PC1	PC2	PC3	PC4
Imputed vs. Reg.	0.248	0.084	0.324	0.412
NIPALS vs. Reg.	0.639	0.463	0.103	0.391

VII. CONCLUSION

The study evaluated the effectiveness of various methods for handling missing data in Principal Component Analysis (PCA) on ecological datasets. The methods investigated included NIPALS, mean imputation, and regularized Singular Value Decomposition (SVD). The results showed that the choice of method significantly impacts the explained variance and loadings, which are crucial for interpreting the underlying data structure.

NIPALS, which can handle missing data directly without prior imputation, degrades with increasing levels of missingness, suggesting it might not be suitable for datasets with substantial missing entries. Mean imputation maintains a high correlation in loadings and low differences in explained variance even with considerable missing- ness but risks introducing bias, especially if the missingness mechanism is not random. Regularized SVD demonstrated the most consistent performance across all levels of missingness, indicating its robustness for handling missing data in PCA.

For ecological datasets, the choice of handling method is critical, with regularized SVD being the most reliable method. NIPALS may not perform well under severe data sparsity, and mean imputation is a simpler alternative but must be used cautiously due to potential biases. Future research should explore alternative imputation techniques and their effects on PCA outcomes, as well as sensitivity analyses to determine optimal regularization parameters for SVD.

Interest Conflicts:

The author(s) declare(s) that there is no conflict of interest concerning the publishing of this paper.

Funding Statement:

No funding for this paper.

VII. REFERENCES

[1] Pech, R., Hao, D., Pan, L., Cheng, H., Zhou, T.: Link prediction via matrix completion. Europhysics Letters 117(3), 38002 (2017)

[2] Panuju, D.R., Paull, D., Griffin, A.L.: Change Detection Techniques Based on Multispectral Images for Investigating Land Cover Dynamics (2020). https://doi. org/10.3390/rs12111781.

[3] Wold, S., Esbensen, K.H., Geladi, P.: Principal component analysis (1987). https://doi.org/10.1016/0169-7439(87)80084-9.

[4] Mehareb, E.M., Gad-Allah, A.: Yield and quality of some sugarcane varieties as affected by irrigation number (2020). https://doi.org/10.21608/svuijas.2020. 38830.1023.

[5] Gui, Y., Barber, R., Ma, C.: Conformalized matrix completion. Advances in Neural Information Processing Systems 36, 4820–4844 (2023)

[6] Butcher, Smith, B.J.: Feature Engineering and Selection: A Practical Approach for Predictive Models (2020). https://doi.org/10.1080/00031305.2020. 1790217. https://doi.org/10.1080/00031305.2020.1790217

[7] Hossie, T.J., Gobin, J., Murray, D.L.: Confronting missing ecological data in the age of pandemic lockdown. Frontiers in Ecology and Evolution 9, 669477 (2021)

[8] Harel, O., Mitchell, E.M., Perkins, N.J., Cole, S.R., Tchetgen, E.J.T., Sun, B., Schisterman, E.F.: Multiple Imputation for Incomplete Data in Epidemiologic Studies (2017). https://doi.org/10.1093/aje/kwx349

[9] Schafer, J.L., Graham, J.W.: Missing data: Our view of the state of the art. (2002). https://doi.org/10.1037/1082-989x.7.2.147.

[10] Enders, C.K.: Applied Missing Data Analysis (2010). http://library.mpib-berlin. mpg.de/toc/z2010 1182.pdf

[11] L- opucki, R., Kiersztyn, A., Pitucha, G., Kitowski, I.: Handling missing data in ecological studies: Ignoring gaps in the dataset can distort the inference. Ecological Modelling 468, 109964 (2022)

[12] Nakagawa, S.: Missing data: mechanisms, methods, and messages. Ecological statistics: Contemporary theory and application, 81–105 (2015)

[13] Taugourdeau, S., Villerd, J., Plantureux, S., Huguenin-Elie, O., Amiaud, B.: Filling the gap in functional trait databases: use of ecological hypotheses to replace missing data. Ecology and Evolution 4(7), 944–958 (2014)

[14] Johnson, T.F., Isaac, N.J., Paviolo, A., Gonz´alez-Su´arez, M.: Handling missing values in trait data. Global Ecology and Biogeography 30(1), 51–62 (2021)

[15] Xiao, J., Bulut, O.: Evaluating the performances of missing data handling methods in ability estimation from sparse data. Educational and Psychological Measurement 80(5), 932–954 (2020)

[16] Hadeed, S.J., O'rourke, M.K., Burgess, J.L., Harris, R.B., Canales, R.A.: Imputation methods for addressing missing data in short-term monitoring of air pollutants. Science of the Total Environment 730, 139140 (2020)

[17] Austin, M.P.: Spatial prediction of species distribution: an interface between ecological theory and statistical modeling. Ecological modelling 157(2-3), 101–118 (2002)

[18] Su, H., Yao, W., Wu, Z., Zheng, P., Du, Q.: Kernel low-rank representation with elastic net for China coastal wetland land cover classification using gf-5 hyperspectral imagery. ISPRS Journal of Photogrammetry and Remote Sensing 171, 238–252 (2021)

[19] Rivera-Mun˜oz, L., Giraldo-Forero, A.F., Martinez-Vargas, J.: Deep matrix factorization models for estimation of missing data in a low-cost sensor network to measure air quality. Ecological Informatics 71, 101775 (2022)

[20] Udell, M., Horn, C., Zadeh, R., Boyd, S., et al.: Generalized low rank models. Foundations and Trends® in Machine Learning 9(1), 1–118 (2016)

[21] Zliobaite, I.: Recommender systems meet species distribution modeling. In: Perspectives@ RecSys (2021)

[22] Zˇliobaite˙, I.: Recommender systems for fossil community distribution modeling. Methods in ecology and evolution 13(8), 1690–1706 (2022) https://doi.org/10. 1111/2041-210x.13916

[23] Chen, Z., Wang, S.: A review on matrix completion for recommender systems. Knowledge and Information Systems 64(1), 1–34 (2022)

[24] Bertsimas, D., Li, M.L.: Fast exact matrix completion: A unified optimization framework for matrix completion. Journal of Machine Learning Research 21(231), 1–43 (2020)

[25] Beattie, J.R., Esmonde-White, F.: Supplementary figures for exploration of principal component analysis: Deriving pca visually using spectra

[26] Seu, K., Kang, M.-S., Lee, H.: An intelligent missing data imputation techniques: A review. JOIV: International Journal on Informatics Visualization 6(1-2), 278– 283 (2022)