Nigerian Journal of Operations Research (NIJOR)

https://nijor.org.ng/
pISSN: 2141-7938
eISSN: 2141-7938

Handling Multicollinearity in Logistic Regression via PCA, ICA, and GPCA: Theory, Simulation and Application

Akinrefon, Adesupo Adeoye[1*], Akerele Josephine Oluwabusayomi[2], Reuben, Benham Zangaluka[3], Ayuba Madu Yami[4]

[1,2,3] Department of Statistics, Faculty of Physical Sciences, Modibbo Adama University, Yola, Nigeria
[4] Department of Operations Research, Faculty of Computing, Modibbo Adama University, Yola, Nigeria

ARTICLE INFO

Article history:
Received 29 June 2026
Revised 7 July 2026
Accepted 9 July 2026
Available online 19 July 2026

Keywords:

Principal Component Analysis, Independent Component Analysis, Generalized Principal Component Analysis, Multicollinearity, Logistic regression, Dimension reduction

ABSTRACT

Multicollinearity poses a serious challenge in logistic regression analysis, leading to inflated standard errors, unstable coefficient estimates, and unreliable statistical inference. This study investigates the comparative performance of three (3) dimension reduction techniques — Principal Component Analysis (PCA), Independent Component Analysis (ICA), and Generalized Principal Component Analysis (GPCA) — as remedies for multicollinearity in binary logistic regression. The theoretical framework establishes the asymptotic properties of maximum likelihood estimators under transformed predictors, including consistency, Fisher information, and asymptotic normality via the Central Limit Theorem. A Monte Carlo simulation study was conducted using eight predictor variables across five sample sizes (n = 50, 100, 500, 1000, 10000) and three levels of inter-predictor correlation (ρ = 0.50, 0.75, 0.95), with 2, 3, and 4 retained components evaluated under each scenario. Model performance was assessed using the Akaike Information Criterion (AIC), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Deviation (MAD). The methods were further validated on a real-life dataset of automobile characteristics from the STATA repository, comprising nine predictor variables exhibiting substantial multicollinearity as confirmed by Variance Inflation Factor (VIF) analysis. Simulation results demonstrate that GPCA consistently achieves superior predictive accuracy over PCA and ICA across varying sample sizes, correlation levels, and numbers of retained components, as evidenced by lower MSE, RMSE, and MAD values. PCA performs competitively under the AIC criterion and is recommended when small samples constrain model complexity. Real-life data analysis corroborates these findings, with GPCA outperforming competing methods on three out of four selection criteria. These results affirm GPCA as the preferred dimension reduction technique for logistic regression in the presence of multicollinearity, with PCA serving as a viable alternative in small-sample setting.

1. Introduction

Multicollinearity is a common problem in statistical regression analysis; it occurs when two or more predictor variables are highly correlated with each other, making it difficult to distinguish their individual effects on the outcome variable. It can lead to several issues, such as inflated standard errors, unstable estimates of the regression coefficients, and difficulty in interpreting the coefficients (Belsley *et al.*, 1980).

[*] *Corresponding author.*
E-mail address: akinrefon.adesupo@mau.edu.ng

Often, handling multicollinearity involves removal of one or more correlated variables from the model, but it is argued that this results in loss of important information, and can also introduce bias in the estimates of the remaining coefficients (Hair *et. al.,* 2010).

Multicollinearity problems also exist in logistic regression analysis, and the use of dimension reduction techniques has been proposed as a solution to handle this issue. Several studies have shown that dimension reduction methods can improve the reliability and accuracy of logistic regression models in the presence of multicollinearity. McInnes *et al.* (2018) demonstrated that PCA can effectively reduce the impact of multicollinearity on logistic regression models, leading to more reliable and accurate predictions. Li *et al.* (2018) used same to reduce the dimensionality of a large set of predictor variables in a study of lung cancer patients, resulting in improved prediction accuracy and model stability. This study seeks to identify the most effective between Principal Component Analysis (PCA), Generalized Principal Component Analysis (GPCA), and Independent Component Analysis (ICA) for handling multicollinearity in logistic regression analyses.

PCA remains the default remedy for multicollinearity in multiple linear and logistic regression because it transforms a set of correlated predictors into a smaller set of mutually orthogonal components while preserving the bulk of total variance. Youssef, *et al*. (2023) in a panel-data context, showed that when near-collinearity among predictors produced unstable fixed-effects estimates, converting predictors into principal components restored well-behaved standard errors and improved out-of-sample prediction relative to ordinary least squares. In a clinical research, Mishra *et al.* (2022) applied PCA ahead of Cox proportional-hazards and logistic regression models built on correlated lipid biomarkers, concluding that PCA-derived covariates yielded more stable hazard and odds ratios than models using the raw, collinear predictors directly. Also, in the health-sciences, Ljubicic, et al (2021) used PCA to condense correlated steroid-hormone measurements in children with congenital adrenal hyperplasia into interpretable composite scores for monitoring treatment adequacy over time. These studies affirm that PCA is computationally inexpensive, interpretable through loadings, and directly compatible with standard generalized linear model software.

Robust and kernel extensions of PCA can be found in researches bothering on its fragility to outliers and heavy-tailed data. For instance, He, *et al*. (2023) proposed a robust PCA procedure based on a characteristic-function transformation of the data; Fayomi *et al.* (2024) replaced the Gaussian likelihood underlying classical PCA with a Cauchy likelihood, thereby producing principal components that are markedly less sensitive to contamination in high-dimensional datasets while remaining competitive with classical PCA when no outliers are present. Avella-Medina, *et al*. (2025) developed theoretical perturbation bounds for kernel PCA and applied the method to characterize the dependence structure of multivariate extremes. PCA is an unsupervised learning technique seeks to maximize the variance of the data along the principal components; it identifies the directions (components) that capture the most variation in the data, when the data has linear relationships between features –that is, expressing one feature as a function of the other(s). PCA compresses data while retaining most information, by choosing just the right number of features (components). Its implicit connection to the Gaussian distribution is undesirable for discrete data such as binary and multi-category responses or counts (Landgraf & Lee, 2019).

ICA is an unsupervised linear dimensionality reduction technique that aims to find a new set of variables that are statistically independent and non-Gaussian (Mocquin, 2023); the goal of ICA is to recover the independent data given from the observation that are linearly dependent on another data. ICA mainly focuses on maximizing the non-linearity by using non-linear transformation function (Nanga *et al.,* 2021). Where PCA seeks orthogonal directions of maximum variance under an implicit Gaussian assumption, ICA seeks a linear unmixing that renders recovered components maximally statistically independent, typically by maximizing the underlying non-Gaussian nature. ICA-derived components are often more directly interpretable than principal components even though both methods yield linear combinations of the observed variables. Li *et al.* (2024) reviewed the theoretical foundations and recent algorithmic advances of ICA while Kouřil *et al.* (2023), applied multivariate ICA to newborn metabolic screening data and found that ICA-derived scores identified affected infants with accuracy comparable to conventional, disease-specific reference ranges. In imaging genetics, Zhang *et al.* (2020), combined canonical correlation analysis with statistical-independence and structural-sparsity constraints closely related to ICA to relate neuroimaging phenotypes to genetic variables. In nutritional metabolomics, Khodorova *et al.* (2021), used ICA-adjacent multivariate decomposition of urinary acyl-carnitine profiles to distinguish signatures of caloric intake from those of dietary protein content; again illustrating the method's comparative advantage when distinct physiological processes are assumed to act as independent, non-Gaussian sources.

"Generalized PCA" most commonly refers to extensions of classical PCA from the Gaussian, squared-loss setting to the broader exponential family of distributions, using Bregman divergences in place of the Euclidean loss implicit in ordinary PCA. Smallman and Artemiou (2022), situated exponential-family PCA within this Bregman-divergence framework, showing that classical PCA is recovered as the Gaussian special case while binary, count, and other non-Gaussian data types admit analogous low-rank representations via appropriately chosen link functions. Nielsen (2024), gave a complementary theoretical grounding by characterizing the divergence structures induced by the cumulant and partition functions of exponential families, clarifying the information-geometric relationships that underpin generalized PCA's loss functions. Nababan, et al (2024), used PCA as a dimensionality-reduction preprocessing step ahead of a custom multiclass logistic regression classifier applied to imbalanced medical datasets, reporting that PCA-conditioned inputs improved classification accuracy, precision, and recall relative to models using raw, correlated clinical features.

A distinct sense of "generalized PCA" concerns dimension reduction for matrix- or tensor-valued time series, where classical vector PCA does not directly apply. He *et al.* (2024) proposed Generalized PCA estimators for large-dimensional matrix factor models that explicitly accommodate heteroscedastic noise across rows, columns, and time, deriving asymptotic distributions for the resulting factor estimates and adaptive thresholding procedures for the associated covariance structure. Agyekum *et al.* (2024), examined how sample size interacts with multicollinearity in high-dimensional logistic regression, recommending a minimum sample size of approximately 500 observations to maintain stable multicollinearity diagnostics, a finding directly relevant to PCA- and GPCA-based preprocessing decisions in moderately sized clinical or survey datasets typical of epidemiological research.

GPCA is designed to handle various types of data using the generalized linear model framework. In contrast to the existing approach of matrix factorizations for exponential family data, GPCA provides low-rank estimates of the natural parameters by projecting the saturated model parameters. This difference in formulation leads to the favorable properties that the number of parameters does not grow with the sample size and simple matrix multiplication suffices for computation of the principal component scores on new data. (Landgraf & Lee, 2019)

Evidence from literature reveal that the PCA and ICA differ fundamentally in objective, while PCA maximizes variance under an implicit second-order, Gaussian-consistent criterion, ICA maximizes statistical independence through non-Gaussian assumption. Although both procedures are linear transformations of the same data. Also, robust and kernelized PCA variants (He *et al.*, 2023; Fayomi *et al.*, 2024; Avella-Medina *et al.*, 2025) address PCA's sensitivity to outliers and nonlinear structure without abandoning its variance-based objective. It is realized from the review that generalized PCA in the Bregman-divergence, exponential-family sense (Smallman & Artemiou, 2022; Nielsen, 2024) explicitly changes the loss function to match the distributional nature of the data, offering a more principled alternative to applying ordinary PCA to binary, count, or compositional data, although it comes at the cost of greater computational and interpretive complexity.

Overall, PCA, ICA, and generalized PCA constitute a set of dimensionality-reduction strategies with the goal of representing correlated, high-dimensional data through a smaller set of derived components, but distinguished by their underlying statistical assumptions.

2. Methodology

Let $Y_i \in \{0,1\}$ and $X_i \in \mathbb{R}^p$. The logistic model is given as $Y_i \sim Bernoulli(\pi_i)$

$$\pi_i = P(Y_i = 1|X_i) = \frac{\exp(X_i^T\beta)}{1+\exp(X_i^T\beta)} \quad (1)$$

with log-likelihood of $\ell(\beta) = \sum_{i=1}^{n}[y_i X_i^T\beta - \log(1+e^{X_i^T\beta})]$
When X^TX ill-conditioned, estimation is becomes unreliable.

2.1 Dimension Reduction Methods

Let $Z = XW$, where $W \in R^{pxk}$; where W contains eigenvectors of the covariance matrix;

a) The Principal Component Analysis (PCA) solves

$$\max_w w^T \sum w \quad subject\ to: ||w|| = 1 \quad (2)$$

b) The Independent Component Analysis (ICA) assumes

$$X = AS \quad (3);$$

where S are independent components; A is the mixing matrix.

c) The Generalized Principal Component Analysis (GPA) solves

$$\max_{rank(L)\leq k} D(\Theta||L) \quad (4)$$

Where $D(.)$ is a Bregman divergence.
Using the transformed predictors, the logistic regression with reduced predictors, is given as

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = Z_i^T\beta \quad (5)$$

with corresponding log-likelihood based on transformed predictors given as

$$\ell(\beta) = \sum_{i=1}^{n}[y_i Z_i^T\beta - \log(1+e^{Z_i^T\beta})] \quad (6)$$

Respectively, the score function, hessian function and the Fisher information are given as follows

$$U_n(\theta) = \sum_{i=1}^{n} Z_i(Y_i - \pi_i)$$

$$H_n(\theta) = -\sum_{i=1}^{n} Z_i Z_i^T \pi_i(1-\pi_i)$$

$$I_n(\theta) = \sum_{i=1}^{n} Z_i Z_i^T \pi_i(1-\pi_i) \quad (7)$$

Asymptotic Properties:

1. Consistency: Under standard regular conditions, the MLE $\hat{\theta}_n$ is consistent if $\hat{\theta}_n \xrightarrow{p} \theta_0$. This follows from concavity of the log-likelihood and the law of large numbers.

2. Asymptotic Normality:

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \rightarrow \mathcal{N}\left(0, \mathrm{I}(\theta_0)^{-1}\right)$$

By Taylor's expansion,

$$0 = U_n(\hat{\theta}) = U_n(\theta_0) + H_n(\tilde{\theta})(\hat{\theta} - \theta_0)$$

Rearranging gives:

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) = -\left(\frac{1}{n} H_n(\tilde{\theta})\right)^{-1} \frac{1}{\sqrt{n}} U_n(\theta_0)$$

By the Central Limit Theorem,

$$\frac{1}{\sqrt{n}} U_n(\theta_0) \rightarrow \mathcal{N}\left(0, \mathrm{I}(\theta_0)\right)$$

and the Law of Large Numbers,

$$\frac{1}{n} H_n(\tilde{\theta}) \rightarrow -I(\theta_0)$$

Thus,

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \rightarrow \mathcal{N}\left(0, \mathrm{I}(\theta_0)^{-1}\right)$$

2.2. Simulation Study

Eight (8) explanatory variables $(X_1, X_2, \ldots X_8)\ where\ X \sim N(0,1)$ were simulated in varying sample sizes of 50, 100, 500, 1000, and 10000; with ρ as the correlation between the explanatory variables in their covariance matrix. The explanatory variables are generated using the simulation procedure stated in (Lukman *et al.*, 2019).

$$x_{ij} = (1 - \rho^2)^{\frac{1}{2}} Z_{ij} + \rho Z_{ij} \qquad i = 1,2, \ldots, n; j = 1, 2, \ldots, 8 \tag{9}$$

Where ρ is the correlation between the explanatory variables. The values of ρ are chosen to be 0.5, 0.75, and 0.95. The outcome variable, $Y_i \sim Bernoulli(\pi_i)$

Where $\pi_i = \frac{e^{x\beta}}{1+e^{x\beta}}$ defining a binary outcome.

The data was subjected to PCA, GPCA and ICA for dimension reduction before the logistic regression model was fit to assess their capacities in handling multicollinearity issues. Also, real life dataset on the status/condition of automobiles with predictor variables such as price, mpg, headroom, gear ratio, length, trunk, weight, turn, and displacement was also used to determine the best dimension reduction technique.

3. Results

This section presents and discusses the results of the Monte Carlo simulation study and the real-life automobile dataset analysis. The three dimension reduction techniques — Principal Component Analysis (PCA), Independent Component Analysis (ICA), and Generalized Principal Component Analysis (GPCA) — are evaluated for their effectiveness in handling multicollinearity within logistic regression. Their comparative performance is assessed using four criteria: the Akaike Information Criterion (AIC), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Deviation (MAD). Lower values across all four criteria indicate a better-fitting and more precise model.

The simulation is designed to reflect a range of practical scenarios, varying both the degree of multicollinearity among predictors and the available sample size. The real-life analysis further grounds these findings in an applied context, allowing for practical recommendations. The simulation covers five sample sizes (n = 50, 100, 500, 1000, and 10,000) and three levels of inter-predictor correlation (ρ = 0.50, 0.75, and 0.95), representing mild, moderate, and severe multicollinearity respectively. Within each scenario, 2, 3, and 4 retained components are examined separately to assess how the number of components influences model performance.

Following the simulation, the methods are applied to a real-life dataset of automobile vehicles sourced from the STATA repository. The dataset contains nine predictor variables — Price, Miles per Gallon (Mpg), Headroom, Trunk, Weight, Length, Turn, Displacement, and Gear Ratio. Using the variable, rep78 – repair record, the outcome variable is taken as either having a good or bad condition record. Summary statistics, inter-variable correlation analysis, and Variance Inflation Factor (VIF) diagnostics are first presented to characterize the nature and extent of the multicollinearity in the real-life data before the dimension reduction methods are applied.

3.1 Discussion of Findings

Table 1 presents simulation results when two (2) components are retained from the reduction process. It assesses the performance of the PCA, ICA and GPCA at 50%, 75% and 95% correlation between eight predictor variables considered using criteria such as AIC, MSE, RMSE and MAD. Given that AIC evaluates model fit with complexity penalty; MSE/RMSE evaluate predictive error, while MAD evaluates robustness.

At a sample size of 50 and across all correlation levels, PCA is selected as yielding better model fitness given the AIC value of 61.004 while GPCA is selected as suited for model prediction accuracy and robustness based on the MSE, RMSE, and MAD criteria. This pattern is also observed when the multicollinearity level goes up to 75% and 95% respectively. The GPCA is adjudged best in dimension reduction given these stated conditions. Generally, across all scenarios, GPCA appears superior in predictive stability, PCA sometimes provides more parsimonious models, while ICA occasionally performs competitively under moderate collinearity. Overall, GPCA seems to have more consensus across the sample sizes and across the levels of collinearity when retaining two (2) components but PCA and ICA also show some good performance based on AIC criterion and at lower collinearity levels.

When three (3) components are retained as shown on Table 2, all the criteria suggest that GPCA is best since their values are the least amidst the lot, at sample size of 50 with 50% correlation. Suggesting that it is the model best fit as the number of components increase and also yields lowest predictive error margin and improved model robustness. At higher collinearity levels, model fitness juggles between ICA and PCA when sample size is 50. Higher sample size when collinearity is at 50%, ICA and GPCA seem to perform favorably across criteria but PCA maintains stable model fit as collinearity rises. Generally, GPCA maintained best predictive accuracy and model robustness while PCA model fit is consistent. ICA also shows promise under moderate collinearity. Examining the performance of the models at the listed sample sizes and collinearity level when four (4) components are retained, the results on Table 3, shows that PCA performance falls behind ICA and GPCA under model appropriateness and predictive ability and robustness. GPCA performs consistently well under model fit and predictive error margin based on their MSE, MAD and AIC.

Table 1: Selecting 2 components using PCA, ICA and GPCA at 50%, 75% and 95% correlation in 50, 100, 500, 1000 and 10000 sample sizes

		Multicollinearity								
		50%			**75%**			**95%**		
Sample Size	**Performance Measures**	**PCA**	**ICA**	**GPCA**	**PCA**	**ICA**	**GPCA**	**PCA**	**ICA**	**GPCA**
50	AIC	**61.00415**	61.37239	67.54742	**63.19263**	63.44948	70.27467	**63.61601**	63.62238	70.12604
	MSE	0.189874	0.189906	**0.17942**	0.201638	0.202983	**0.188948**	0.201121	0.20122	**0.186546**
	RMSE	0.435745	0.435745	**0.42358**	0.449041	0.449041	**0.434681**	0.448466	0.448466	**0.431909**
	MAD	0.374919	0.376582	**0.35187**	0.397603	0.400084	**0.37299**	0.398373	0.39849	**0.369976**
100	AIC	137.5301	137.6888	**136.9136**	127.1504	**126.6837**	127.2643	131.9388	**131.2409**	138.4397
	MSE	0.233275	0.23382	**0.209249**	0.209311	0.208514	**0.187724**	0.220619	0.219026	**0.213346**
	RMSE	0.482985	0.482985	**0.457438**	0.457505	0.457505	**0.433271**	0.469701	0.469701	**0.461894**
	MAD	0.466109	0.466992	**0.419244**	0.418392	0.416547	**0.375749**	0.440628	0.437504	**0.425582**
500	AIC	**661.6061**	662.8293	666.545	629.2283	**609.2839**	632.6099	626.5861	**626.1475**	631.6098
	MSE	0.232265	0.232818	**0.229987**	0.217302	**0.208597**	0.21457	0.216304	0.216092	**0.214108**
	RMSE	0.481939	0.481939	**0.47957**	0.466156	0.466156	**0.463217**	0.465085	0.465085	**0.462718**
	MAD	0.464135	0.465245	**0.459519**	0.434512	**0.41705**	0.428907	0.432045	0.431636	**0.427573**
1000	AIC	**1314.497**	1317.153	1322.553	**1298.446**	1298.551	1298.514	**1208.602**	1210.628	1216.013
	MSE	0.231257	0.23187	**0.230813**	0.227636	0.227624	**0.225358**	0.207477	0.207871	**0.2069**
	RMSE	0.480891	0.480891	**0.48043**	0.477112	0.477112	**0.474719**	0.455496	0.455496	**0.454862**
	MAD	0.462441	0.463679	**0.461555**	0.455091	0.4551	**0.450464**	0.41484	0.415656	**0.413646**
10000	AIC	**11362.94**	11566.53	11368.53	12717.51	**12693.03**	12723.58	**12261.88**	12268.78	12267.44
	MSE	0.193526	0.197719	**0.193445**	0.222679	**0.222128**	0.222612	0.212564	0.212711	**0.21246**
	RMSE	0.439916	0.439916	**0.439823**	0.471889	0.471889	**0.471818**	0.461047	0.461047	**0.460933**
	MAD	0.386716	0.395164	**0.386552**	0.445317	**0.44421**	0.445178	0.425019	0.425316	**0.424821**

Table 2: Selecting 3 components using PCA, ICA and GPCA at 50%, 75% and 95% correlation in 50, 100, 500, 1000 and 10000 sample sizes

		Multicollinearity								
		50%			**75%**			**95%**		
Sample Size	**Performance measures**	**PCA**	**ICA**	**GPCA**	**PCA**	**ICA**	**GPCA**	**PCA**	**ICA**	**GPCA**
50	AIC	75.01471	75.372	**71.73491**	**63.3122**	63.91481	70.47707	69.39758	**69.35628**	74.06376
	MSE	0.238092	0.239981	**0.167059**	0.184779	0.187267	**0.162562**	0.214241	0.21411	**0.182718**
	RMSE	0.487946	0.487946	**0.408729**	0.42986	0.42986	**0.403189**	0.462862	0.462862	**0.427455**
	MAD	0.476771	0.4804	**0.33301**	0.371412	0.376522	**0.321444**	0.427404	0.427095	**0.359019**
100	AIC	**113.6129**	113.9396	122.9117	129.975	**129.5627**	125.8098	131.1652	**130.4116**	141.5658
	MSE	0.17942	0.180413	**0.172245**	0.21097	0.210686	**0.174929**	0.214104	0.212412	**0.206375**
	RMSE	0.42358	0.42358	**0.415024**	0.459315	0.459315	**0.418245**	0.462714	0.462714	**0.454285**
	MAD	0.356122	0.357798	**0.341457**	0.421931	0.420902	**0.347334**	0.427956	0.424607	**0.412592**
500	AIC	652.8165	**646.0182**	664.821	**621.1386**	621.4314	628.93	**626.1101**	626.1823	626.6158
	MSE	0.226697	**0.223596**	0.225781	0.213589	0.213745	**0.211332**	0.214837	0.214949	**0.208173**
	RMSE	0.476127	0.476127	**0.475164**	0.462157	0.462157	**0.459708**	0.463505	0.463505	**0.456259**
	MAD	0.453512	**0.447294**	0.451703	0.42634	0.42663	**0.421424**	0.429551	0.429722	**0.416726**
1000	AIC	1268.668	**1265.333**	1276.875	**1204.134**	1237.228	1212.855	**1204.852**	1206.681	1212.173
	MSE	0.220424	**0.219016**	0.219049	0.206911	0.212941	**0.206212**	0.205293	0.20567	**0.203894**
	RMSE	0.469493	0.469493	**0.468027**	0.454875	0.454875	**0.454106**	0.453093	0.453093	**0.451546**
	MAD	0.440713	0.438532	**0.438075**	0.412986	0.426218	**0.411213**	0.411552	0.412331	**0.408567**
10000	AIC	**11428.93**	11630.06	11433.08	12558.41	**12522.13**	12565.16	12301.36	**12299.45**	12306.51
	MSE	0.194995	0.199204	**0.194792**	0.219199	**0.2184**	0.219061	0.21342	0.213379	**0.213217**
	RMSE	0.441582	0.441582	**0.441352**	0.468187	0.468187	**0.468039**	0.461974	0.461974	**0.461754**

Table 3: Selecting 4 components using PCA, ICA and GPCA at 50%, 75% and 95% correlation in 50, 100, 500, 1000 and 10000 sample sizes

		Multicollinearity								
		50%			**75%**			**95%**		
Sample size	**Performance measures**	**PCA**	**ICA**	**GPCA**	**PCA**	**ICA**	**GPCA**	**PCA**	**ICA**	**GPCA**
50	AIC	39.7356	40.31771	**28.0000**	59.60842	60.47709	**53.59065**	63.10772	**62.76311**	67.26288
	MSE	0.098505	0.10027	**0.0s248**	0.168516	0.17138	**0.078383**	0.181146	0.179663	**0.120849**
	RMSE	0.313855	0.313855	**0.0498**	0.410507	0.410507	**0.27997**	0.425613	0.425613	**0.347634**
	MAD	0.19173	0.195243	**0.0175**	0.331602	0.338075	**0.15965**	0.359705	0.356838	**0.249514**
100	AIC	112.85	**110.1678**	124.0859	108.8814	**107.8804**	113.3062	**135.2498**	135.4008	139.091
	MSE	0.172213	0.16773	**0.164567**	0.160277	0.158679	**0.141023**	0.218661	0.218864	**0.185005**
	RMSE	0.414985	0.414985	**0.405668**	0.400347	0.400347	**0.375531**	0.467612	0.467612	**0.430122**
	MAD	0.344246	0.334391	**0.32385**	0.32381	0.320038	**0.279952**	0.437059	0.437596	**0.373419**
500	AIC	554.9863	**548.5979**	562.5018	**609.4114**	610.8849	611.7181	**637.9375**	638.4946	649.2291
	MSE	0.184913	0.182011	**0.180358**	0.207166	0.207908	**0.201129**	0.219816	0.220024	**0.217008**
	RMSE	0.430015	0.430015	**0.424686**	0.455155	0.455155	**0.448474**	0.468846	0.468846	**0.465841**
	MAD	0.368588	0.363133	**0.359882**	0.413676	0.415051	**0.40148**	0.43906	0.439511	**0.43337**
1000	AIC	1259.26	**1224.717**	1269.772	**1202.248**	1212.391	1209.428	**1251.241**	1251.789	1260.702
	MSE	0.217574	**0.209777**	0.21611	0.204519	0.206664	**0.202154**	0.216142	0.216195	**0.214034**
	RMSE	0.466449	0.466449	**0.464876**	0.452238	0.452238	**0.449616**	0.464911	0.464911	**0.462638**
	MAD	0.435271	**0.419866**	0.432165	0.409862	0.414192	**0.405255**	0.432058	0.432229	**0.427961**
10000	AIC	12346.72	**12114.66**	12356.7	**12431.48**	12493.15	12442.06	**12107.37**	12108.48	12113.64
	MSE	0.21438	**0.209171**	0.214235	0.216057	0.21739	**0.215897**	0.208644	0.208666	**0.208409**
	RMSE	0.463011	0.463011	**0.462855**	0.464819	0.464819	**0.464647**	0.456776	0.456776	**0.456518**
	MAD	0.428631	**0.418296**	0.428328	0.432202	0.434908	**0.431879**	0.417622	0.417668	**0.417119**

Table 3 presents simulation results when four (4) components are retained from the reduction process. It assesses the performance of the PCA, ICA and GPCA at 50%, 75% and 95% correlation between eight predictor variables considered using criteria such as AIC, MSE, RMSE and MAD.

At a sample size of 50, GPCA gave the least values of AIC, MSE, RMSE, and MAD at all the different levels of collinearity except at 95% collinearity where ICA has the lowest AIC. The GPCA is adjudged best in dimension reduction given these stated conditions.

At a sample size of 100, the ICA is chosen when the collinearity level is at 50% with AIC of 110.1678, MSE of 0.164567, RMSE of 0.405668 and MAD of 0.32385. At 75% collinearity level, AIC criterion selects ICA while at 95%, PCA has the lowest AIC. GPCA has the lowest MSE, RMSE and MAD. GPCA gave the least values on these criteria as the best dimension reduction technique.

At a sample size of 500, GPCA is still selected based on the MSE (0.180358), RMSE (0.424686), and MAD (0.359882) while the AIC criterion choses ICA (609.4114) as the better option at 50% collinearity. At 75% and 95% collinearity, PCA has the lowest AIC while GPCA has the smallest MSE, RMSE and MAD values.

In a sample of 1000, GPCA is selected based on RMSE (0.464876), while ICA has the lowest AIC (1224.717), MSE (0.209777), and MAD (0.419866) when there is about 50% collinearity; at 75% collinearity, PCA has the smallest AIC, while GPCA has the least MSE, RMSE and MAD values. While at 95% collinearity, PCA is selected based on the AIC criterion while the MSE, RMSE and MAD criteria suggest the use of GPCA.

In a sample of 10000, GPCA is selected based on RMSE (0.462855), while ICA has the lowest AIC (12114.66), MSE (0.209171), and MAD (0.418296) when there is about 50% collinearity; at 75% and 95% collinearity, PCA has the smallest AIC, while GPCA has the least MSE, RMSE and MAD values.

Overall, GPCA seems to have more consensus across the sample sizes and across the levels of collinearity when retaining four (4) components but PCA and ICA also show some good performance based on AIC criterion and at lower collinearity levels.

3.2 Analysis on Real Life dataset

Table 4: Summary statistics of the predictor variables

Variable	Mean	Std. Dev.	min	max
Price	6165.257	2949.496	3291.000	15906
Mpg	21.297	5.786	12.000	41
Headroom	2.993	.846	1.500	5
Trunk	13.757	4.277	5.000	23
Weight	3019.459	777.194	1760.000	4840
Length	187.932	22.266	142.000	233
Turn	39.649	4.399	31.000	51
Displacement	197.297	91.837	79.000	425
Gear Ratio	3.015	.456	2.190	3.89

Table 5: showing values the strength of relationship between predictor variables

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1) Price	1.000								
(2) Mpg	-0.469	1.000							
(3) Headroom	0.115	-0.414	1.000						
(4) Trunk	0.314	-0.582	0.662	1.000					
(5) Weight	0.539	-0.807	0.483	0.672	1.000				
(6) Length	0.432	-0.796	0.516	0.727	0.946	1.000			
(7) Turn	0.310	-0.719	0.424	0.601	0.857	0.864	1.000		
(8) Displacement	0.495	-0.706	0.474	0.609	0.895	0.835	0.777	1.000	
(9) Gear_Ratio	-0.314	0.616	-0.378	-0.509	-0.759	-0.696	-0.676	-0.829	1.000

Table 4 presents summary statistics (mean, standard deviation, minimum and maximum values), on the nine (9) predictor variable considered in this dataset. It can be observed that Price has a mean of 6165.257 and a standard deviation of 2949.496 while Mpg has a mean of 21.297 and a standard deviation of 5.786. The means and standard deviation of other predictor variables as well as their minimum and maximum values can also be observed from Table 4. Next, the correlation between these variables is examined and presented on Table 5.

From Table 5, it can be seen that there are significant levels of relationship between the nine variables. While some show correlation values less than 50%, some others show strength of correlation that are 50% and above with respect to other predictor variables. Thus, we explore the status of the variance inflation factor (VIF), and their tolerance limits.

Table 6: Values Variance Inflation Factor (VIF)

Factor	Price	Mpg	Headroom	Trunk	Weight	Length	Turn	Displacement	Gear Ratio
VIF	1.966069	2.8833	1.965894	2.743544	**18.98122**	**15.09589**	4.098218	**5.410579**	3.20737
Tolerance	0.508629	0.3468	0.508674	0.364492	0.052684	0.066243	0.244008	0.184823103	0.311782

From the Table 6, Price or current value of the automobiles has a VIF of 1.966069 and a corresponding tolerance limit of 0.508629. Since its VIF is less than 5, Price is retained in the final model analysis; next, Mpg has a VIF of 2.8833 and tolerance limit of 0.3468. It is also retained in the final analysis. Headroom, trunk, turn and Gear-ration have VIF values of 1.965894, 2.743544, 4.098218, and 3.20737 respectively. Since these are below 5, they are also retained in the model. But Weight, Length, and Displacement show significant evidence of high multicollinearity with some other predictor variable as such pose the problem of multicollinearity. It is noticeable that their tolerance limits are close to zero. In fitting Logistic regression to the dataset on automobile status, we first carry out the dimension reduction procedure using the PCA, ICA, and the GPCA.

Table 7: Fitting Logistic Regression on Initial dataset (Raw Dataset)

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	6.7462432	8.1306162	0.83	0.407
Price	-0.0002275	0.0001418	-1.604	0.109
Mpg	-0.1160734	0.0909004	-1.277	0.202
Headroom	-0.0431177	0.4620687	-0.093	0.926
Trunk	-0.018743	0.1074213	-0.174	0.861
Weight	0.0018306	0.0016798	1.09	0.276
Length	-0.0931968	0.0499374	-1.866	0.062
Turn	0.1978803	0.1427277	1.386	0.166
Displacement	0.0088603	0.007979	1.11	0.267
Gear_Ratio	0.1687258	1.1523574	0.146	0.884
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				
Null deviance:	99.099	On	73 d.f	
Residual deviance:	77.047	On	64 d.f	
AIC: 97.047				

Table 8: Selection criteria on the reduced dataset using the three methods

	AIC	MSE	RMSE	MAD
PCA	96.3296	0.183234	0.428058	0.370248
ICA	**93.07989**	0.172637	0.428058	0.3513
GPCA	98.44	**0.123588**	**0.35155**	**0.250701**

From Table 8, it is clearly observed that GPCA satisfies three (3) out of the four criteria. Hence, we fit the logistic regression on the reduced components from the GPCA.

Table 9: Fitting logistic regression on the reduced dataset

Coefficients:	Estimate St	d. Error	z value	Pr(>|z|)	
(Intercept)	20.0189	13.6732	1.464	0.1432	
contrib.Dim.1	5.366	2.8651	1.873	**0.0611**	.
contrib.Dim.2	2.0266	1.0568	1.918	**0.0551**	.
contrib.Dim.3	1.7316	0.791	2.189	**0.0286**	*
contrib.Dim.4	0.2056	0.9688	0.212	0.8319	
contrib.Dim.5	-0.3434	0.4222	-0.813	0.416	
contrib.Dim.6	0.8515	0.6182	1.377	0.1684	
Signif. Codes	: 0 '***' 0	.001 '**	' 0.01 '	*' 0.05 '	.' 0.1 ' ' 1
	(Dispersion parameter taken to be 1				
Null Deviance	99.099	73 df			
Residual deviance	58.44	on 54			
AIC: 98.44					

4. Conclusions:

The results of this study consistently demonstrate that GPCA is a better dimension reduction technique for addressing multicollinearity in logistic regression, owing to its grounding in the generalized linear model framework which aligns naturally with binary response data. Unlike PCA and ICA, which derive components by maximizing variance or statistical independence under Gaussian assumptions, GPCA operates directly on the natural parameters of the exponential family distribution, making it inherently more appropriate for discrete outcomes such as binary logistic regression. This methodological alignment translates into consistently lower predictive error across diverse simulation conditions and on real-life data. PCA, while inferior on MSE, RMSE, and MAD, performs well under the AIC criterion and is a computationally simpler alternative that may be preferred in small-sample settings ($n \leq 100$) where overfitting risk is elevated and the marginal gain from GPCA may not justify its added complexity. ICA, though conceptually distinct in seeking statistically independent rather than uncorrelated components, does not offer a systematic advantage over PCA or GPCA in the logistic regression context examined here; its performance is competitive at lower collinearity levels but does not generalize as reliably across the full range of conditions studied. Practitioners dealing with moderate to severe multicollinearity ($\rho \geq 0.75$) and moderate to large sample sizes ($n \geq 100$) are therefore strongly advised to adopt GPCA as the dimension reduction method of choice before fitting logistic regression models. For small samples or mild multicollinearity, PCA remains a robust and interpretable alternative. Future research should extend these comparisons to multinomial and ordinal logistic regression settings, explore the integration of GPCA with regularization approaches such as ridge and LASSO penalties, and examine performance under non-normal predictor distributions and missing data scenarios.

Author Contributions

Conceptualization: A.A.A. and A.J.O.; Methodology, A.A.A. and A.J.O.; Software: A.A.A.; Validation: R.B.Z. and A.M.Y; Formal analysis: A.J.O.; Resources: A.M.Y.; Data curation: A.A.A.; Writing—original draft preparation: A.A.A. and A.J.O.; Writing—review and editing: R.B.Z.; A.J.O. and A.Y.M. All authors have read and agreed to the final version of the manuscript.

Funding

This research received no external funding.

Data Availability Statement

A real-life datasets on automobile characteristics from the STATA repository, comprising nine predictor variables was used.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research was not funded by any grant.

References:

Agyekum, G. O., Adarkwa, S. A., & Kusi, R. Y. (2024). Impact of sample size on multicollinearity with high dimensional data in logistic regression analysis. *International Journal of Innovation and Development*, 1(3). Retrieved from https://ijid.kstu.edu.gh/index.php/ijid/article/view/26

Avella-Medina, M., Davis, R. A., & Samorodnitsky, G. (2025). Insights into kernel PCA with application to multivariate extremes. *SIAM Journal on Mathematics of Data Science*, 7(2), 777–801. https://doi.org/10.1137/24M1678635

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). Regression diagnostics: Identifying influential data and sources of collinearity. John Wiley & Sons.

Fayomi, A., Pantazis, Y., Tsagris, M., & Wood, A. T. (2024). Cauchy robust principal component analysis with applications to high-dimensional data sets: A. Fayomi et al. *Statistics and Computing*, *34*(1), 26. https://doi.org/10.1007/s11222-023-10328-x

Hair Jr, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). Multivariate data analysis. In *Multivariate data analysis* (pp. 785-785).

He, L., Yang, Y., & Zhang, B. (2023). Robust PCA for high-dimensional data based on characteristic transformation. *Australian & New Zealand Journal of Statistics*, 65(2), 127–151. https://doi.org/10.1111/anzs.12385

He, Y., Hou, Y., Liu, H., & Wang, Y. (2024). Generalized principal component analysis for large-dimensional matrix factor model (Working paper). Retrieved from https://arxiv.org/abs/2411.06423

Khodorova, N. V., Rietman, A., Rutledge, D. N., Schwarz, J., Piedcoq, J., Pilard, S., Siebelink, E., Kok, F. J., Tomé, D., & Azzout-Marniche, D. (2021). Urinary medium-chained acyl-carnitines sign high caloric intake whereas short-chained acyl-carnitines sign high-protein diet within a high-fat, hypercaloric diet in a randomized crossover design dietary trial. *Nutrients*, 13(4), 1191. https://doi.org/10.3390/nu13041191

Kouřil, Š., de Sousa, J., Fačevicová, K., Gardlo, A., Muehlmann, C., Nordhausen, K., Friedecký, D., & Adam, T. (2023). Multivariate independent component analysis identifies patients in newborn screening equally to adjusted reference ranges. *International Journal of Neonatal Screening*, 9(4), 60. https://doi.org/10.3390/ijns9040060

Landgraf, A. J., & Lee, Y. (2020). Generalized principal component analysis: Projection of saturated model parameters. *Technometrics*, *62*(4), 459-472.http://doi.org/10.1080/00401706.2019.1668854

Li, Y., Zhang, M., Bian, X., Tian, L., & Tang, C. (2024). Progress of independent component analysis and its recent application in spectroscopy quantitative analysis. *Microchemical Journal*, 202, 110836. https://doi.org/10.1016/j.microc.2024.110836

Li, C., Zhao, S., Shi, J., & Wang, X. (2018). Principal component regression analysis for lung cancer patients based on microarray data. *Oncology Letters*, 15(2), 2665-2672.

Ljubicic, M. L., Madsen, A., Juul, A., Almstrup, K., & Johannsen, T. H. (2021). The application of principal component analysis on clinical and biochemical parameters exemplified in children with congenital adrenal hyperplasia. *Frontiers in Endocrinology*, 12, 652888. https://doi.org/10.3389/fendo.2021.652888

Lukman A, Ayinde K, Binuomote S, Onate C.A. & Clement A (2019). Modified Ridge-type estimator to combat multicollinearity: Application to chemical data. *Journal of Chemometrics* https://doi.org/10.1002/cem.3125

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*

Mishra, A., Nair, N. S., Kumar, K. H., Binu, V. S., & Heesh, S. S. (2022). Application of Principal Component Analysis in Dealing with Multicollinearity in Modelling Clinical Data. *Journal of Clinical & Diagnostic Research*, *16*(7), 15. https://doi.org/10.7860/JCDR/2022/55379.16629

Nababan, A. A., Sutarman, Zarlis, M., & Nababan, E. B. (2024). Multiclass logistic regression classification with PCA for imbalanced medical datasets. *Mathematical Modelling of Engineering Problems*, 11(9), 2377–2387. https://doi.org/10.18280/mmep.110911

Nanga, S., Bawah, A., Acquaye, B., Billa, M., Baeta, F., Odai, N., Obeng, S. and Nsiah, A. (2021) Review of Dimension Reduction Methods. *Journal of Data Analysis and Information Processing*, 9, 189-231. http://doi.org/10.4236/jdaip.2021.93013.

Nielsen, F. (2024). Divergences induced by the cumulant and partition functions of exponential families and their deformations induced by comparative convexity. *Entropy*, 26(3), 193. https://doi.org/10.3390/e26030193

Odjesa, E., Okonofua, F. E., Bolarinwa, O. A., Zhou, J., & Sawangdee, Y. (2024). An empirical analysis of the demand for family planning satisfied by modern methods among married or in-union women in Nigeria: Application of multilevel binomial logistic modelling technique. *PLOS ONE*, 19(3), e0300744. https://doi.org/10.1371/journal.pone.0300744

Smallman, L., & Artemiou, A. (2022). A literature review of (sparse) exponential family PCA. *Journal of Statistical Theory and Practice*, 16, 14. https://doi.org/10.1007/s42519-021-00238-4

Youssef, A. H., Mohamed, E. S., & Abdel Latif, S. H. (2023). Handling multi-collinearity using principal component analysis with the panel data model. *EUREKA: Physics and Engineering*, (1), 177–188. https://doi.org/10.21303/2461-4262.2023.002582

Zhang, Y., Peng, P., Ju, Y., Li, G., Calhoun, V. D., & Wang, Y.-P. (2020). Canonical correlation of imaging genetics data based on statistical independence and structural sparsity. *IEEE Journal of Biomedical and Health Informatics*, 24(9), 2621–2629. https://doi.org/10.1109/JBHI.2020.2972580