Logistic regression and discriminant analysis of academic staff-mix by rank via research performance in a university setting

V. U. Ekhosuehi and A. Iduseri[1]

Department of Statistics, University of Benin, P.M.B. 1154, Benin City, Nigeria

To safeguard the university system, guidelines for appointments and promotions of academic staff are designed to serve as a benchmark for assessing and appraising the staff. Appraisal of academic staff in Nigerian universities has become a subject of controversy in recent times. This paper is aimed at finding a classifier for the discrimination of academic staff of a university system of known states or staff categories, viz.: S_1 – *Lecturer I and below,* S_2 – *Senior Lecturer and* S_3 – *Professor/Associate Professor, into latent subgroups on the basis of their research proficiency. A combination of cluster analysis and linear discriminant analysis was first used as a framework for identifying three latent subgroups, viz., mover, mediocre and stayer, with the quantity of scholarly publications, quality of academic journals in the system publish in, and the author-level citation index as input variables. Principal component analysis in combination with logistic regression was used to investigate and classify a (training) data set of a cross-section of academics of several categories with diverse research features from different universities within Nigeria. The results revealed that there are more stayers in* S_2, *and more movers in* S_3. *A comparison of the staff categories indicates that the research performance of academics in* S_3 *outstrips those in* S_1 *and* S_2, *and that academics in* S_1 *did better than those in* S_2. *The methods reported here have potential utility for the latent intra-class categorisation of staff of the research-oriented system within the mover-mediocre-stayer paradigm. The method is useful for shortlisting applicants for interview to a more appropriate staff category of the system.*

Keywords: academic discriminant analysis; cluster analysis; logistic regression; mover-mediocre-stayer paradigm principal; component analysis

1. Introduction

Universities are citadels of knowledge-producing institutions and repositories of knowledge for advancement. Academics in the university system advance and communicate knowledge through their scholarly publications and this core duty is vital for their career growth in the academia. In addition, the quantity and quality of scholarly publications indexed by international bibliographic databases of authors affiliated to a university provide useful information on the bibliometric strength of the university (Demir, 2018). Universities in Nigeria are regulated by the National Universities Commission (NUC). This Commission has a mandate to set standards and assure quality in university education in Nigeria. This mandate is achieved through periodic evaluation of programmes and the framing of benchmark guidelines for appointments and promotions for the system. To be considered for promotion, the academic staff must have attained competency in terms of publications in reputable journals and been at a grade for a prescribed minimum duration (usually three years). Nonetheless, universities also have the prerogative to improve on the guidelines in line with their visions and missions. The

[1] Corresponding Author; Email: augustine.iduseri@uniben.edu

guidelines are a point-based system on a set of criteria which include: academic qualifications, publications and creative works, teaching/professional experience, conferences, administrative experience and general contribution. In many instances a high relative weight is assigned to publications and creative works, especially for promotions/appointments to the rank of Senior Lecturer and beyond. Therefore, it can be said that the quantity and quality of scholarly publications are a precondition for advancement in the academic reward structure for an individual aspiring to the senior academic ranks.

According to the hierarchical nomenclature for academic staff-mix by rank of the NUC, the states (staff categories) of the university system can be defined as follows: S_1 – Lecturer I and below, S_2 – Senior Lecturer and S_3 – Professor/Associate Professor (Ekhosuehi and Omosigho, 2018). These staff categories play a major role in determining the statutory staffing requirements for programme accreditation in the university system. To safeguard the university system, guidelines for appointments and promotions (of academics) are designed to serve as a benchmark for assessing and appraising the staff. Appraisal of academics in Nigerian universities has become a subject of controversy in recent times. Some are of the opinion that academics in Nigeria are rewarded for quantity of publications, especially in foreign journals, rather than the quality of the publications and the journals in which such articles are published in (see Adomi and Mordi (2003) and the references cited therein). As already reported, Nigeria is ranked second (next to India) among countries that publish in predatory/fake journals. Academics who do this do not want to experience the rigour and trauma of review and rejection because they lack appropriate research skills, but want to achieve a high score based on the number of publications and get promoted (Demir, 2018). As a result, some academics in Nigeria are rather seen as possessing the research proficiency that is appropriate for their grade level. Research proficiency here is measured by the quantity of published articles in reputable journals indexed by international databases (such as SSCI/SCI/AHCI, Scopus, etc.) and the author-level citation index (for example, h-index). Insofar the author is aware of only one university in Nigeria (University of Ilorin, Ilorin), where any academic curriculum vitae (CV) is not countenanced for including publication(s) in predatory journal(s).

Udom and Ebedoro (2019) split the states of a manpower system (which includes the university manpower system) into three sub-categories according to the latent sources of heterogeneity, viz. mover, mediocre and stayer. The "movers" are characterised by high promotion probabilities and high career growth; the "mediocre" are characterised by intermediate promotion probabilities and average career growth; and the "stayers" have the lowest promotion probabilities within the groups. Since scholarly publication is a universal precondition for academic advancement in the university system (Adomi and Mordi, 2003), the mover, mediocre and stayer are akin to high, average and low research proficiency, respectively. Harris and Kaine (1994) explore the relationship between individual productivity in research and their preferences and perceptions about research-oriented issues. The latent intra-category subgroups proposed by Udom and Ebedoro (2019) and the subclasses of Harris and Kaine (1994) are similar when applied to a research-oriented system such as a university, the only difference is in nomenclature.

More related to this study are the determinants of research performance by Harris and Kaine (1994). Their focus was on the calculation of publication points per capita and the extent to which lecturers have differential access to research support facilities which in turn would inform their preferences and perceptions about research-oriented issues. Hence, this study is quite different from that of Harris and Kaine (1994) in terms of their thematic focus. Our focus is on research proficiency which is a more appropriate latent source of heterogeneity because it is a function of individual motivation (Harris and Kaine, 1994). It is an index of visibility and advancement in the academic reward structure. Assessing and classifying academics on the basis of research proficiency can reduce the risk of appointing and promoting less competent personnel, especially to the higher ranks where supervision of postgraduate students are required. This suggests a need to revisit the extant guidelines on appointments and promotions.

This study is aimed at finding a classifier for the discrimination of academic staff of a university system into latent subgroups on the basis of their research proficiency. More specifically, the study proposes a technique-dependent separation rule within the framework of discriminant analysis. The study models the chances of classifying a staff using principal component analysis (PCA) in combination with logistic regression (LR). In this sense, this study contributes to the literature on modelling heterogeneity in a manpower research-oriented system.

2. Preliminaries on Learning Algorithms used in this Study

To make this paper clearer to a wider audience, a 'snapshot' on discriminant analysis, logistic regression and principal component analysis is provided. Linear discriminant analysis (discriminant analysis or DA) involves the use of a training set of observations of known class membership to generate functions that separate these observations into the specified classes optimally and to determine how to allocate new observations into groups (Gaynanova and Wang, 2019 and Hardle and Simar, 2007). Thus, the classical linear discriminant analysis (LDA) is a supervised dimensionality reduction method for finding the directions that maximally separate the different classes while minimising the spread within classes. The performance of the classical DA is poor when the classes have non normal-like or multi-modal mixture distributions or when the within-class covariance matrix is nearly singular, especially in high-dimensional data (Gao and Li, 2019). There are a number of areas in which discriminant analysis can be applied. Examples include: sex determination (Renjith *et al*., 2019), wheat productivity (Chauhan *et al*., 2020), prediction of academic achievement (Cornell-Farrow and Garrard, 2020), image set classification (Gao and Li, 2019), fault classification in industrial processes (Lim *et al*., 2018; Zheng *et al*., 2019), identification of seed variety (Xia *et al*., 2019), identification of oil pollutants (Deming *et al*., 2018), forensic classification of black inkjet prints (Oravec *et al*., 2019), cross-view classification in computer vision(You *et al*., 2019), chemogenomics of the bioactivity of proteins (Tavernier *et al*., 2019), classification of emerging drugs (Setser and Smith, 2018), diagnosis of autism spectrum disorders (Zou *et al*., 2018), differentiation of breast lesion (Boudaghi and Saen, 2014), prediction of flood susceptibility (Choubin *et al*., 2019), etc. In spite of the wide applicability of discriminant analysis, it has received little attention in the manpower

planning literature (De Feyter, 2006). The reasons not being that discriminant analysis is not applicable to manpower data, but could be that manpower planners are not so familiar with the use of discriminant analysis as a screening technique for achieving a more homogeneous personnel category, which is key to building Markov manpower models (Bartholomew *et al*., 1991). This study, therefore contributes to manpower literature. Real life applications of DA models have shown that some covariates could be relevant, redundant or independent for the analysis (Maugis *et al*., 2011). These ideas have served the motive for feature construction methods such as feature selection and feature extraction (Gao and Li, 2019). However, the aforementioned challenge can be resolved by applying a notable dimensionality reduction technique such as the principal component analysis (PCA).

PCA is a well-known unsupervised feature extraction method that does not require any prior knowledge of the groups (Affes and Hentati-Kaffel, 2019; Orevac *et al.*, 2019; Setser and Smith, 2018). As common to unsupervised methods, the PCA does not perform class discrimination, but focuses on the eigenvalues and eigenvectors of the correlation (or covariance) matrix of the predictor variable vector to represent the original data in a lower dimensional subspace satisfying certain criteria (Nwosu *et al*., 2016; Parmet *et al*., 2010).In other words, the PCA method extracts the most important information from the data set and compresses the data information by keeping only the most important variables that could explain the heterogeneity in the system. This initial dimension reduction approach is also important to do away with redundancy as the raw data may convey information with some degree of redundancy (Maugis *et al*., 2011).

DA method as a generative classifier is valid only when the sample covariance matrices are invertible. This limitation is circumvented in the Logistic Regression (LR) approach to DA. LR is a supervised learning algorithm for estimating the probability of an outcome or class variable. The principle of maximum likelihood is commonly used in fitting the LR model and the coefficients maximising the likelihood of the data are obtained using iterative techniques (for example, the Newton-Raphson algorithm, iterative scaling algorithm, conjugate gradient ascent algorithm and estimation of distribution algorithms). Parameter estimation using the Newton-Raphson algorithm is limited if the training data set is sparse (i.e., the number of variables exceeds the number of observations). Bi and Jeske (2010) show that LR is less deteriorated by class-conditional classification noise (CCC-Noise) compared to linear discriminant analysis (LDA) and that, when the Mahalanobis distance between two multivariate normal populations is small, LR tends to be more tolerable to CCC-Noise compared to LDA. Affes and Hentati-Kaffel (2019) compared the classification and prediction of both LDA and LR models with and without misclassification costs. They found that the logit model outperforms the LDA in terms of correct classification rate. There is no need to further stress the supportive empirical relevance of LR in discriminant analysis as this has been done elsewhere (Cornell-Farrow and Garrard, 2020; Dattalo, 1995; Katos, 2007; Press and Wilson 1978).

This study contributes to the use of bibliometric data on publications and citations to classify academics based on LR and LDA. Early research employed LDA (Harris and Kaine, 1994; Oravec et *al*., 2019; Zhou *et al.*, 2007). The LR model is more appropriate for the

problem at hand because assumptions like normality of the covariates are not required, and it can accommodate all kinds of measurement scale: nominal, ordinal, interval or ratio.

3. Methodology

3.1 Linear discriminant analysis

Classical discriminant analysis focuses on the Gaussian model defined by the density

$$f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right), \quad \mathbf{x} \in R^p, \tag{1}$$

with location parameter vector $\boldsymbol{\mu} \in R^p$ and non-singular covariance matrix $\boldsymbol{\Sigma}$ (Manjunath *et al*., 2012). For a variable vector $\mathbf{x}$ that comes from one of two $p-$dimensional Gaussian populations with class labels 1 and 2, the discriminant function for classifying an observation $\mathbf{x} \in R^p$ between $f_{\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1}(\mathbf{x}|1)$ and $f_{\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2}(\mathbf{x}|2)$ is

$$Df_Q(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) - \log\frac{c_2 p_2 |\boldsymbol{\Sigma}_1|^{1/2}}{c_1 p_1 |\boldsymbol{\Sigma}_2|^{1/2}}, \tag{2}$$

where p_1, p_2 and c_1, c_2 are the corresponding prior probabilities and costs of misclassification to the first and second population respectively. The function Df_Q is quadratic in $\mathbf{x}$. If the populations differ in mean, but not in covariance, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, and the discriminant function Df_Q is linear in $\mathbf{x}$. The observation vector is classified to class 1 if the inequality $Df_Q(\mathbf{x}) \geq 0$ is fulfilled.

Arevalillo and Navarro (2012) establish theoretical foundations regarding the classification of multivariate data in experiments where the multivariate normality of the observations must be discarded as a realistic assumption. They suggested the elliptically contoured distributions as the alternative when the normality assumption fails. A variable vector $\mathbf{x} \in R^p$ is said to follow an elliptically contoured distribution if its density function is given by

$$f(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Lambda},g) = c_p|\boldsymbol{\Lambda}|^{-1/2} g((\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Lambda}^{-1}(\mathbf{x}-\boldsymbol{\mu})), \tag{3}$$

where $\boldsymbol{\mu} \in R^p$, $\boldsymbol{\Lambda}$ is a $p \times p$ positive definite matrix, $c_p = \dfrac{\Gamma(p/2)}{\pi^{p/2}\int_0^\infty t^{p/2-1}g(t)dt}$, and g is a non-negative real-valued function such that $\int_0^\infty t^{p/2-1}g(t)\,dt < \infty$.

3.2 Principal component analysis

This study proposes the use of the PCA method to achieve an uncorrelated blend of covariates before applying discriminant analysis. Let $\{\mathbf{x}_i\}_{i=1}^n$ be a data matrix constructed from the variable vector $\mathbf{X} = (X_1, X_2, \cdots, X_p)$ and $\boldsymbol{\Gamma}$ the $p \times p$ correlation matrix for the observations set $\{\mathbf{x}_i\}_{i=1}^n$. It is needless to transform the observation set because the components of $\boldsymbol{\Gamma}$ are scale-invariant (Rencher, 2002). Let $\boldsymbol{\Gamma} = \mathbf{EVE}'$ be the spectral decomposition of $\boldsymbol{\Gamma}$, where $\mathbf{V} = diag(\lambda_1, \lambda_2, \cdots, \lambda_p)$ is a matrix of the eigenvalues

of $\mathbf{\Gamma}$ and $\mathbf{E}$ is a matrix whose columns correspond to the eigenvectors of the eigenvalues of $\mathbf{\Gamma}$. The principal components, P_d, are obtained by searching for a dimension along which the observations $\{\mathbf{x}_i\}_{i=1}^n$ are maximally separated as a linear combination of the variables in $\mathbf{X}$ as

$$P_d = \sum_{r=1}^{p} a_{dr} X_r, \quad d = 1,2,\cdots,m < p, \tag{4}$$

where a_{dr} is an entry in $\mathbf{E}$ of $\mathbf{\Gamma}$ corresponding to the eigenvalue λ_d and m is the number of components retained using certain criteria (for instance, the Kaiser-Guttman (KG) rule of one or the 80% rule, etc. Ekhosuehi (2017), Nwosu *et al.* (2020) and Parmet *et al.* (2010)). The component scores are computed from the expression

$$(p_{id})_{n\times m} = \{\mathbf{x}_i\}_{i=1}^n \mathbf{\Psi}, \tag{5}$$

where p_{id} is the score for the d th component of the ith observation and $\mathbf{\Psi}$ is a $p \times m$ matrix of the columns of $\mathbf{E}$ corresponding to the eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_m$. With PCA, the factor scores are used as inputs in the development of the LR classifier.

3.3 Logistic regression

This study relies on the success story on logistic regression as a supervised classification technique to construct the classifier for the three-group intra-category classification problem. According to the academic staff-mix strata by rank defined by NUC (Ekhosuehi and Omosigho, 2018), there are three hierarchical categories for the system: S_1 – Lecturer I and below, S_2 – Senior Lecturer and S_3 – Professor/Associate Professor. These categories are the possible outputs of the response variable. Thus, the response variable is ordinal. The generalised linear model for ordinal responses is ordinal logistic regression. For a discussion on a general class of regression models for ordinal data, see McCullagh (1980).

Define δ_{ij} to be an indicator variable, which is created from the training data set $\{(\mathbf{x}_i, g_i)\}_{i=1}^n \subset R^p \times S$ as follows: for $j = j^*$,

$$\delta_{ij^*} = \begin{cases} 1 & \text{if} \quad g_i = S_{j^*} \\ 0 & \text{otherwise} \end{cases}.$$

In what follows, let R^m be the feature subspace constructed by applying the principal component method on $\mathbf{X} \in R^p$ and let $\phi_j(\mathbf{x}_i)$ be the probability of classifying a staff in category j with a score vector $\mathbf{x}_i = (p_{i1}, p_{i2}, \cdots, p_{im}) \in R^m$ as being in that same category. Then the logit model is defined to be

$$\log\left(\frac{\phi_1(\mathbf{x}_i)}{\phi_2(\mathbf{x}_i) + \phi_3(\mathbf{x}_i)}\right) = \beta_{01} + \sum_{d=1}^{m} \beta_d p_{id}, \tag{6}$$

and

$$\log\left(\frac{\phi_1(\mathbf{x}_i) + \phi_2(\mathbf{x}_i)}{\phi_3(\mathbf{x}_i)}\right) = \beta_{02} + \sum_{d=1}^{m} \beta_d p_{id}, \tag{7}$$

where the intercepts β_{01} and β_{02} depend on the category and the β_d, $d = 1,2,3,\cdots,m$, are the regression coefficients. These parameters are usually estimated from data by means of the maximum likelihood method. This is achieved by maximising the log-likelihood function L with respect to the parameter vector $\boldsymbol{\beta} = (\beta_{01}, \beta_{02}, \beta_1, \cdots, \beta_m)$, where

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(\delta_{i1}\left(\beta_{01} + \sum_{d=1}^{m}\beta_d p_{id} \right) - (\delta_{i1}+\delta_{i2})\log\left(1+\exp\left(\beta_{01} + \sum_{d=1}^{m}\beta_d p_{id} \right)\right)\right.$$
$$+ \delta_{i2}\log\left(\exp\left(\beta_{02} + \sum_{d=1}^{m}\beta_d p_{id} \right) - \exp\left(\beta_{02} + \sum_{d=1}^{m}\beta_d p_{id} \right)\right)$$
$$\left. - (\delta_{i2}+\delta_{i3})\log\left(1+\exp\left(\beta_{02} + \sum_{d=1}^{m}\beta_d p_{id} \right)\right)\right). \tag{8}$$

The following system of $m+2$ equations with $m+2$ parameters has to be solved:

$$\frac{\partial L}{\partial \beta_{01}} = \sum_{i=1}^{n}\left(\left(\delta_{i1} - \delta_{i2}\frac{\phi_1(\mathbf{x}_i)}{\phi_2(\mathbf{x}_i)} \right)(1-\phi_1(\mathbf{x}_i)) \right) = 0, \tag{9}$$

$$\frac{\partial L}{\partial \beta_{02}} = \sum_{i=1}^{n}\left(\left(\delta_{i2}\frac{\phi_2(\mathbf{x}_i)+\phi_3(\mathbf{x}_i)}{\phi_2(\mathbf{x}_i)} - (\delta_{i2}+\delta_{i3}) \right)(\phi_1(\mathbf{x}_i)+\phi_2(\mathbf{x}_i)) \right) = 0, \tag{10}$$

and

$$\frac{\partial L}{\partial \beta_d} = \sum_{i=1}^{n} p_{id}\left((\delta_{i1}+\delta_{i2})(1-\phi_1(\mathbf{x}_i)) - (\delta_{i2}+\delta_{i3})(\phi_1(\mathbf{x}_i)+\phi_2(\mathbf{x}_i)) \right) = 0, \tag{11}$$

$d = 1,2,\cdots,m$. There is no analytical solution to this system of nonlinear equations. This study resorts to using the Newton-Raphson numerical method as it has a rapid convergence rate with a reasonable choice of starting estimates for the parameter vector, $\hat{\boldsymbol{\beta}}^{(0)}$ (Minka, 2004). In this regard, new estimates are obtained for $\hat{\boldsymbol{\beta}}^{(\zeta)}$, $\zeta = 1,2,\cdots$, using an updating formula given as follows

$$\hat{\boldsymbol{\beta}}^{(\zeta)} = \hat{\boldsymbol{\beta}}^{(\zeta-1)} - \mathbf{H}^{-1}(\hat{\boldsymbol{\beta}}^{(\zeta-1)})\mathbf{U}^{(\zeta-1)}, \ \zeta = 1,2,\cdots, \tag{12}$$

where the transpose $\mathbf{U}' = \left(\frac{\partial L}{\partial \beta_{01}} \quad \frac{\partial L}{\partial \beta_{02}} \quad \frac{\partial L}{\partial \beta_1} \quad \frac{\partial L}{\partial \beta_2} \quad \cdots \quad \frac{\partial L}{\partial \beta_m} \right)$ is the $1\times(m+2)$ score vector, ζ indicates the number of iterations and $\mathbf{H}^{-1}(\hat{\boldsymbol{\beta}}^{(\zeta-1)})$ is the inverse of the Hessian matrix, $\mathbf{H}(\hat{\boldsymbol{\beta}}^{(\zeta-1)})$, at the $(\zeta-1)$th iteration. The Hessian matrix is given by

$$\mathbf{H}(\hat{\boldsymbol{\beta}}) = \left[\frac{\partial^2 L}{\partial \hat{\boldsymbol{\beta}} \partial \hat{\boldsymbol{\beta}}'} \right]_{(m+2)\times(m+2)}, \tag{13}$$

where the matrix elements are given by

$$\left.\frac{\partial^2 L}{\partial \beta_{01}^2}\right|_{\beta_{01}=\hat{\beta}_{01}} = -\sum_{i=1}^{n}\left(\left(\delta_{i1} + \delta_{i2}\left(1 + \frac{(\phi_1(\mathbf{x}_i)+\phi_2(\mathbf{x}_i))\phi_3(\mathbf{x}_i)}{\phi_2^2(\mathbf{x}_i)} \right)\right)\phi_1(\mathbf{x}_i)(\phi_2(\mathbf{x}_i)+\phi_3(\mathbf{x}_i)) \right), \tag{14}$$

$$\left.\frac{\partial^2 L}{\partial \beta_{01}\partial \beta_{02}}\right|_{\beta_{01}=\hat{\beta}_{01},\beta_{02}=\hat{\beta}_{02}} = \sum_{i=1}^{n}\delta_{i2}\phi_1(\mathbf{x}_i)\phi_3(\mathbf{x}_i)(\phi_1(\mathbf{x}_i)+\phi_2(\mathbf{x}_i))\left(\frac{\phi_2(\mathbf{x}_i)+\phi_3(\mathbf{x}_i)}{\phi_2^2(\mathbf{x}_i)} \right), \tag{15}$$

$$\left.\frac{\partial^2 L}{\partial\beta_{01}\partial\beta_{d+1}}\right|_{\beta_{01}=\hat{\beta}_{01},\,\beta_d=\hat{\beta}_d} = -\sum_{i=1}^{n} p_{id}(\delta_{i1}+\delta_{i2})\phi_1(\mathbf{x}_i)(\phi_2(\mathbf{x}_i)+\phi_3(\mathbf{x}_i)),\ d=1,2,\cdots,m, \tag{16}$$

$$\left.\frac{\partial^2 L}{\partial\beta_{02}^2}\right|_{\beta_{02}=\hat{\beta}_{02}} = -\sum_{i=1}^{n}\left(\left(\delta_{i3}+\delta_{i2}\left(1+\frac{(\phi_2(\mathbf{x}_i)+\phi_3(\mathbf{x}_i))\phi_1(\mathbf{x}_i)}{\phi_2^2(\mathbf{x}_i)}\right)\right)\phi_3(\mathbf{x}_i)(\phi_1(\mathbf{x}_i)+\phi_2(\mathbf{x}_i))\right), \tag{17}$$

$$\left.\frac{\partial^2 L}{\partial\beta_{02}\partial\beta_d}\right|_{\beta_{02}=\hat{\beta}_{02},\,\beta_d=\hat{\beta}_d} = -\sum_{i=1}^{n} p_{id}(\delta_{i2}+\delta_{i3})\phi_3(\mathbf{x}_i)(\phi_1(\mathbf{x}_i)+\phi_2(\mathbf{x}_i)),\ d=1,2,\cdots,m, \tag{18}$$

$$\left.\frac{\partial^2 L}{\partial\beta_d^2}\right|_{\beta_d=\hat{\beta}_d} = -\sum_{i=1}^{n} p_{id}^2\big((\delta_{i1}+\delta_{i2})\phi_1(\mathbf{x}_i)(\phi_2(\mathbf{x}_i)+\phi_3(\mathbf{x}_i))+ (\delta_{i2}+\delta_{i3})\phi_3(\mathbf{x}_i)(\phi_1(\mathbf{x}_i)+\phi_2(\mathbf{x}_i))\big),\ d=1,2,\cdots,m, \tag{19}$$

$$\left.\frac{\partial^2 L}{\partial\beta_s\partial\beta_d}\right|_{\beta_s=\hat{\beta}_s,\,\beta_d=\hat{\beta}_d} = -\sum_{i=1}^{n} p_{is}p_{id}\big((\delta_{i1}+\delta_{i2})\phi_1(\mathbf{x}_i)(\phi_2(\mathbf{x}_i)+\phi_3(\mathbf{x}_i))+ (\delta_{i2}+\delta_{i3})\phi_3(\mathbf{x}_i)(\phi_1(\mathbf{x}_i)+\phi_2(\mathbf{x}_i))\big),\ s\neq d,\ s=1,2,\cdots,m. \tag{20}$$

The iteration is stopped when negligible changes between $\hat{\boldsymbol{\beta}}^{(q)}$ and $\hat{\boldsymbol{\beta}}^{(q-1)}$ are achieved.

3.4 Classification rule

Let $\hat{\phi}_j(\mathbf{x}_i)$ be the estimated class probability. For $j=j^*$, define $\theta_{j^*}(\mathbf{x}_i)=\max_{\forall j}\left(\hat{\phi}_j(\mathbf{x}_i)\right)$. Then a member, $\mathbf{1}(g_i)$, with class label $g_i=S_{j^*}$ of the system is classified into one of the latent subcategories, M_{j^*q}, $q=1,2,3$, as follows.

Case 1 ($j^*=1$). This study uses the cut-off probability of 0.5. The standard/common cut off value of 0.5 was adopted since the three staff categories sizes are equivalently equal which makes the three types of misclassification equally costly. The classification rule is: (i) If $\theta_1(\mathbf{x}_i)=\hat{\phi}_1(\mathbf{x}_i)\geq 0.5$, then $\mathbf{1}(g_i)\rightarrow M_{11}$. (ii) If $\theta_1(\mathbf{x}_i)=\hat{\phi}_1(\mathbf{x}_i)<0.5$, then $\mathbf{1}(g_i)\rightarrow M_{12}$.(iii) Otherwise $\theta_1(\mathbf{x}_i)\neq\hat{\phi}_1(\mathbf{x}_i)$, $\mathbf{1}(g_i)\rightarrow M_{13}$. The arrow $\rightarrow$ means 'is classified as'.

Case 2 ($j^*=2$). (i) If $\theta_2(\mathbf{x}_i)=\hat{\phi}_1(\mathbf{x}_i)$, then $\mathbf{1}(g_i)\rightarrow M_{21}$. (ii) If $\theta_2(\mathbf{x}_i)=\hat{\phi}_2(\mathbf{x}_i)$, then $\mathbf{1}(g_i)\rightarrow M_{22}$. (iii) If $\theta_2(\mathbf{x}_i)=\hat{\phi}_3(\mathbf{x}_i)$, then $\mathbf{1}(g_i)\rightarrow M_{23}$.

Case 3 ($j^*=3$). (i) If $\theta_3(\mathbf{x}_i)=\hat{\phi}_1(\mathbf{x}_i)$, then $\mathbf{1}(g_i)\rightarrow M_{31}$. (ii) If $\theta_3(\mathbf{x}_i)=\hat{\phi}_2(\mathbf{x}_i)$, then $\mathbf{1}(g_i)\rightarrow M_{32}$. (iii) If $\theta_3(\mathbf{x}_i)=\hat{\phi}_3(\mathbf{x}_i)$, then $\mathbf{1}(g_i)\rightarrow M_{33}$.

Using this classification rule, an attempt is made to answer the question: For an applicant having the relevant cognate experience with attribute $\mathbf{x}\in R^p$, what category in S should he/she be shortlisted for interview? Suppose that there is a vacancy in category j. Then the employer should shortlist the available applicants with $\theta_{j^*}(\mathbf{x}_i)$, $j^*\geq j$, for interview.

3.5 The problem setup

Consider a university system with the population of academic staff stratified into k exclusive groups represented by the set $S=\{S_1,S_2,\cdots,S_k\}$, where the entries $S_j\in S$, $j=1,2,\cdots,k$, follow a natural order $S_1<S_2<\cdots<S_k$. These groups create discrimination

among the population. There exists a partition of each category $S_j \in S$ into three subcategories M_{jq}, $q = 1,2,3$, according to the mover-mediocre-stayer paradigm (Udom and Ebedoro, 2019) such that: $S_j = \bigcup_{q=1}^{3} M_{jq}$ and $\bigcap_{q=1}^{3} M_{jq} = \phi$, where M_{j1} is the subcategory of stayers in S_j, M_{j2} is the subcategory of mediocre in S_j, and M_{j3} is the subcategory of movers in S_j. These partitions form the latent classes of intra-category heterogeneity of the system.

Given a (training) data set of size n: $\{(\mathbf{x}_i, g_i)\}_{i=1}^{n} = \{(\mathbf{x}_1, g_1), (\mathbf{x}_2, g_2), \cdots, (\mathbf{x}_n, g_n)\}$, taking values in $R^p \times S$, and drawn from a cross-section of academic staff population in the system, where for each $i \in \{1,2,\cdots,n\}$, $\mathbf{x}_i \in R^p$ is the feature vector and $g_i \in S$ is the class label according to the hierarchical nomenclature for the staff categories. R^p is the $p-$dimensional Euclidean space. Each feature vector $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$ is an observed value of a variable vector $\mathbf{X} \in R^p$, where $\mathbf{X}$ is composed of p random variables $\mathbf{X} = (X_1, X_2, \cdots, X_p)$ with X_r, $r = 1,2,\cdots,p$, being a one-dimensional random variable. The covariates could be number of scholarly publications, number of articles published by quartiles, the author-level metric, length of service or research-oriented qualifications (e.g. Doctor of Philosophy, Doctor of Science, etc.). A decision maker is shown the data set and after evaluating the observations, must decide on the latent subcategory to assign each observation. For an applicant having the relevant cognate experience with attribute $\mathbf{x} \in R^p$, what category in S should he/she be shortlisted for interview?

4. Analysis and Results

For illustrative purposes, this study utilises cross-sectional (training) data on academic staff selected from eighteen (18) universities (including federal, state and private universities) in Nigeria, which comprises 27, 21, 27 staff for category S_1, S_2, S_3, respectively, as the benchmark classification-oriented data set (see Appendix). The use of cross-sectional data is expected to maintain national outlook on the use of the benchmark guidelines on appointments and promotions. The individual academic-level data with details on publications, h-index, position and affiliation are obtained from both ResearchGate (www.researchgate.net) and Google Scholar (scholar.google.com) platforms. These platforms are believed to contain accurate information on the research outputs and grade levels of the selected academics. Nonetheless, academics that their grade levels could not be ascertained where not included in the study. The ResearchGate (RG) score for each author was also collated. The RG score, which is based on bibliographic data on publications, questions, answers and the number of followers on the platform is a measure that indicates how all of one's research is received by peers. For this reason, the RG scores are used for checking the results of the LR classifier. Information on the author-level h-index is obtained from Google Scholar. The h-index is used as a measure of the author-level citations. The information on the journal quartiles is retrieved from Scimago Journal Ranking (SJR). The SJR is a standard index to classify

journals into groups (Q1, Q2, Q3, Q4) according to their impact factors (IFs). Publications not contained in the SJR list, which include books, book chapters, conference proceedings and journal articles, are represented by 'others. The use of SJR is expected to reduce arbitrariness and subjectivity in the ranking of journals. The journal quartiles, other scholarly publications not contained in the SJR list and the h-index are used as the research proficiency covariates.

However, one should not attach undue importance to the data set as the purpose of this illustration is not to solve the national university problem, but to give an insight on the potential utility of the LR model for the latent intra-class categorisation of university academic staff within the mover-mediocre-stayer paradigm. To bring the national problem to practice, a comprehensive database of academics in Nigerian universities and their research features need to be evaluated. As at the time of writing this paper, the author cannot find such database of academics in Nigerian universities.

A combination of cluster analysis and linear discriminant analysis (LDA) was used as a framework for identifying the three latent subgroups (also referred to as low, average and high publication proficiency groups). Zhou *et al.* (2007) also employ cluster analysis as an unsupervised method to divide the target population of authors into smaller groups according to their productivity in terms of mean impact factor, annually citation per publication, first author and corresponding author, before using LDA to confirm the cluster analysis results. Regardless of the kind of university, this study addresses the problem at hand using the following steps.

Table 1: Correlation coefficients between the covariates

	Q1	Q2	Q3	Q4	Others
Q2	(0.783) (0.000)				
Q3	(0.783) (0.000)	(0.783) (0.000)			
Q4	(0.783) (0.007)	(0.783) (0.000)	(0.783) (0.000)		
Others	(0.783) (0.003)	(0.783) (0.000)	(0.783) (0.001)	(0.783) (0.006)	
h-index	(0.783) (0.000)	(0.000) (0.000)	(0.000) (0.000)	(0.000) (0.000)	

Cell contents: Pearson correlation and P-Value

Step 1 Score calculation based on the PCA method

Step 2 Estimate the LR model using the scores

Step 3 Predict the ranks using the calibrated LR model

Step 4 Compare the predictions and the initial classification to assign a latent subgroup to the individual academic.

Step 5 Validate the latent intra-category classification on the basis of the discrimination by the RG score.

Table 1 presents the correlation coefficients between the covariates. It can be seen that all the covariates are positively correlated and that Q1, Q2, Q3 are highly correlated with each other and Q3 is highly correlated with h-index. This is an indication of a strong link between these variables and this is symptomatic of some degree of redundancy in the information they convey. Thus, it is possible to group the variables. In order to ascertain the possibility of reducing the dimension of the data set, this correlation matrix is compared to the identity matrix using the Bartlett's test of sphericity which was significant $\left(\chi^2 = 296.836,\ p = 0.000\right)$, indicating that the sample correlation matrix did not come from a population where the correlation matrix is an identity matrix and that dimension reduction of the data set is valid.

The spectral decomposition of correlation coefficients matrix is carried out to get the eigenvalues and eigenvectors.

$$\mathbf{E} = \begin{pmatrix} 0.440 & 0.388 & 0.083 & -0.157 & 0.764 & 0.202 \\ 0.466 & 0.137 & 0.054 & -0.406 & -0.567 & 0.525 \\ 0.475 & 0.225 & -0.065 & -0.156 & -0.199 & -0.810 \\ 0.300 & -0.591 & -0.718 & -0.120 & 0.168 & 0.051 \\ 0.289 & -0.654 & 0.685 & -0.049 & 0.103 & -0.083 \\ 0.436 & 0.056 & -0.032 & 0.877 & -0.132 & 0.137 \end{pmatrix}.$$

Each column of the matrix, $\mathbf{E}$, contains the eigenvectors of the correlation matrix for the dataset. The first column of $\mathbf{E}$ is the first eigenvector and it is the weights used in the linear combination of the original data in the first principal component.

The eigenvalues are given in Table 2. On the basis of the 80% rule, Table 2 shows that 89.3% of the variation in the dataset is explained by the first three principal components. The most important principal component explains 64.1% of the total variation.

Table 2: Eigen Analysis of the Correlation Matrix

Eigenvalue	3.8483	0.8185	0.6933	0.3370	0.1958	0.1071
Proportion	0.641	0.136	0.116	0.056	0.033	0.018
Cumulative	0.641	0.778	0.893	0.950	0.982	1.000

To calibrate the LR model, the principal component scores are determined using Equation (5). Then the maximum likelihood estimation, which entails the use of the Newton-Raphson iterative scheme is employed. The maximum likelihood estimates, after successive approximations, are obtained at the point where the Euclidean norm between the parameters of the previous and the current iteration is 1.3282×10^{-11}. The results and the measures of association are given in Tables 3 and 4, respectively. For the final approximation, the covariance matrix of the parameters, which is minus the inverse of the Hessian of Equation (13), is

$$-\mathbf{H}^{-1}(\hat{\boldsymbol{\beta}}) = \begin{pmatrix} 0.1956 & 0.1907 & -0.0101 & 0.0054 & 0.0025 \\ 0.1907 & 0.2790 & -0.0144 & 0.0064 & 0.0044 \\ -0.0101 & -0.0144 & 0.0024 & -0.0002 & -0.0013 \\ 0.0054 & 0.0064 & -0.0002 & 0.0067 & 0.0062 \\ 0.0025 & 0.0044 & -0.0013 & 0.0062 & 0.0070 \end{pmatrix}.$$

Table 3: Logistic Regression Results

Parameters	$\hat{\beta}_{01}$	$\hat{\beta}_{02}$	$\hat{\beta}_{1}$	$\hat{\beta}_{2}$	$\hat{\beta}_{3}$
Estimates	0.8734	2.4046	-0.1905	0.0370	0.0812
p-value	0.044	0.000	0.000	0.45	0.97

Test that all slopes are zero: G = 25.91, DF = 3, p-value = 0.000

Table 4: Measures of Association

Pairs	Number	Percent (%)	Summary Measures	
Concordant	1421	76.3	Somers' D	0.53
Discordant	433	23.2	Goodman-Kruskal Gamma	0.53
Ties	9	0.5	Kendall's Tau-a	0.36
Total	1863	100.0		

Deviance = 137.89.

In Table 3, judging by the p-values which measure the evidence against the null hypothesis, the intercept terms and the first principal component are significantly different from zero at the significance level of 5%. This implies that the calibrated LR model has discriminative power and that it could be used for predictive purposes. Instead of the choice of principal components by the 80% rule, the analysis would be as well valid by employing the KG rule of one. The zero p-value for the test that all slopes are zero provides stronger evidence that the relationship between the response variable and the predictors is statistically significant. A cursory look at Table 4 shows that the percentage of concordant pairs is very high. This supports the evidence that the model is statistically significant. Additionally, the summary provided by Somers' D, Goodman-Kruskal Gamma and Kendall's Tau-a indicate a moderate positive strength of association (agreement) between the response and predicted probabilities using the principal components.

Since the system has three categories ($k = 3$) and 75 observations ($n = 75$) were sampled, the degrees of freedom for the estimates with three principal components ($m = 3$) are $df = n(k-1)-(k-1+m) = 145$.If the LR model is an exact fit of the data, the deviance should approximately have the distribution $\chi^2(145)$. But the calculated value of the deviance is not too close to the upper 5% point of $\chi^2(145)$, which is 174.101. This suggests that the model does not fit the data exactly well. This is not a reason to worry since classification is the goal of modelling, the goodness-of-fit of the LR model (model calibration) is no longer of great importance, rather it is the model discrimination (or prediction) between the classes.

Table 5: Predictions and classification to the mover-mediocre-stayer subgroup

Index	$\phi_1(\mathbf{x}_i)$	$\phi_2(\mathbf{x}_i)$	$\phi_3(\mathbf{x}_i)$	LR S_j	RG S_j	Initial S_j	Latent subgroup by LR
1	**0.5842**	0.2824	0.1334	1	1	2	Stayer
2	0.0379	0.1161	**0.8460**	3	3	3	Mover
3	**0.5053**	0.3200	0.1747	1	1	1	Stayer
4	0.2782	**0.3624**	0.3594	2	2	3	Mediocre
5	**0.6514**	0.2449	0.1038	1	1	2	Stayer
6	0.1214	0.2684	**0.6102**	3	3	3	Mover
7	**0.4992**	0.3225	0.1783	1	2	3	Stayer
8	**0.6135**	0.2666	0.1199	1	1	1	Stayer
9	**0.5778**	0.2857	0.1364	1	1	2	Stayer
10	0.2454	0.3552	**0.3994**	3	1	2	Mover
11	0.0416	0.1255	**0.8329**	3	3	3	Mover
12	**0.6191**	0.2635	0.1174	1	1	1	Stayer
13	**0.3873**	0.3578	0.2549	1	2	2	Stayer
14	0.1268	0.2750	**0.5982**	3	3	3	Mover
15	**0.5366**	0.3060	0.1574	**1**	**1**	**3**	**Stayer**
16	0.0000	0.0000	**1.0000**	3	3	3	Mover
17	**0.4609**	0.3372	0.2019	1	2	1	Mediocre
18	**0.4907**	0.3260	0.1833	1	1	1	Mediocre
19	**0.4336**	0.3461	0.2203	1	2	2	Stayer
20	0.1598	0.3081	**0.5320**	3	3	2	Mover
21	**0.6214**	0.2622	0.1164	1	1	1	Stayer
22	**0.4759**	0.3318	0.1924	1	1	1	Mediocre
23	0.0925	0.2279	**0.6796**	3	3	3	Mover
24	**0.6132**	0.2668	0.1201	1	1	1	Stayer
25	0.0540	0.1548	**0.7912**	3	3	3	Mover
26	**0.3784**	0.3595	0.2622	1	1	1	Mediocre
27	**0.6351**	0.2544	0.1105	1	1	1	Stayer
28	0.0687	0.1857	**0.7456**	3	3	3	Mover
29	**0.5833**	0.2829	0.1338	1	1	1	Stayer
30	**0.4456**	0.3424	0.2121	1	1	1	Mediocre
31	**0.4302**	0.3471	0.2227	1	1	2	Stayer
32	**0.4578**	0.3383	0.2039	1	3	1	Mediocre
33	0.2314	0.3506	**0.4180**	3	2	3	Mover
34	**0.4674**	0.3349	0.1977	1	2	3	Stayer
35	0.0744	0.1966	**0.7289**	3	3	3	Mover
36	0.1291	0.2776	**0.5934**	3	2	1	Mover
37	0.3296	**0.3649**	0.3055	2	2	2	Mediocre
38	0.0091	0.0315	**0.9594**	3	3	1	Mover
39	0.0621	0.1722	**0.7657**	3	3	3	Mover
40	0.0251	0.0813	**0.8936**	3	3	3	Mover
41	**0.3720**	0.3606	0.2675	1	2	2	Stayer
42	**0.5223**	0.3125	0.1651	1	1	2	Stayer
43	**0.3969**	0.3558	0.2473	1	1	1	Mediocre
44	**0.5263**	0.3108	0.1629	1	2	1	Stayer
45	**0.3752**	0.3600	0.2648	1	3	1	Mediocre
46	0.0699	0.1880	**0.7421**	3	3	2	Mover
47	**0.3955**	0.3561	0.2484	**1**	**1**	**3**	**Stayer**
48	0.0329	0.1029	**0.8642**	3	3	3	Mover

Table 5 (Cont'd)

49	**0.6391**	0.2521	0.1089	1	1	1	Stayer

50	**0.6272**	0.2589	0.1139	1	2	1	Stayer
51	**0.5379**	0.3054	0.1567	1	1	1	Stayer
52	0.1557	0.3045	**0.5398**	3	2	1	Mover
53	**0.5058**	0.3198	0.1745	1	2	2	Stayer
54	**0.3656**	0.3615	0.2729	1	1	1	Mediocre
55	**0.6078**	0.2697	0.1224	1	1	1	Stayer
56	**0.3727**	0.3604	0.2668	1	3	1	Mediocre
57	0.1141	0.2592	**0.6266**	3	3	2	Mover
58	0.0535	0.1537	**0.7928**	3	3	3	Mover
59	0.1904	0.3305	**0.4791**	3	3	3	Mover
60	**0.4216**	0.3496	0.2288	1	2	2	Stayer
61	**0.4199**	0.3501	0.2301	1	2	2	Stayer
62	**0.4595**	0.3377	0.2028	1	2	3	Stayer
63	0.2707	0.3611	**0.3682**	3	1	3	Mover
64	**0.6125**	0.2671	0.1204	1	1	2	Stayer
65	**0.6221**	0.2618	0.1161	1	1	1	Stayer
66	0.1294	0.2779	**0.5927**	3	3	3	Mover
67	**0.5569**	0.2963	0.1468	**1**	**1**	**3**	**Stayer**
68	**0.5469**	0.3011	0.1519	1	1	2	Stayer
69	0.3320	**0.3648**	0.3032	2	2	2	Mediocre
70	**0.5218**	0.3128	0.1654	1	1	1	Stayer
71	0.0572	0.1618	**0.7811**	3	3	3	Mover
72	0.3154	**0.3651**	0.3195	2	2	2	Mediocre
73	**0.3871**	0.3578	0.2550	1	1	2	Stayer
74	**0.5747**	0.2874	0.1380	**1**	**1**	**3**	**Stayer**
75	0.0246	0.0798	**0.8956**	3	3	3	Mover

The probability values of $\theta_{j*}(\mathbf{x}_i)$ *are made bold in Table 5.*

The calibrated LR model is used to predict the group membership and the results are compared with predictions based on the RG score. Splitting the data set into training and validation data using any splitting criteria will not achieve reasonable precision in the validation, and is likely to produce misleading statistics of the parameters in the targeted population. Therefore, the reason for adopting the use of the RG score as a measure of precision is more realistic in the current context. Table 5 shows the predicted outcomes of the LR model and the discriminant analysis using the RG score. The intra-category classification into the latent subgroups of mover, mediocre and stayer are also contained in Table 5. Statistics show that the LR model is as good as the discriminant analysis using the RG score at 73.33% prediction harmony. Indeed, the LR model using research proficiency covariates can be viewed as an alternative to the discriminant analysis using the RG score. On the grounds of parsimony, the results from the LR model are preferred because the RG score is based on variables that are not relevant to appraising academic staff in Nigeria such as framing of questions, providing answers and number of followers on the RG platform. Moreover, additional predictors can be easily introduced to the LR model if there is a strong reason to do so (e.g., the percentage of contribution of the individual author).

Categorising academic staff into the mover-mediocre-stayer subgroups is intuitive and this indicates the crop of academics in the selection. The categorisation shows the potential for the LR model to predict the latent intra-category outcomes should more detailed information becomes available in future. According to the decision rule for latent

intra-category outcomes the proportion of mover-mediocre-stayer by category is calculated. This is contained in Table 6.

Table 6: Proportion of Mover-Mediocre-Stayer by Category

S_1			S_2			S_3		
M_{11}	M_{12}	M_{13}	M_{21}	M_{22}	M_{23}	M_{31}	M_{32}	M_{33}
48.15%	40.74%	11.11%	66.66%	14.29%	19.05%	25.93%	3.70%	70.37%

The distribution of academics according to the mover-mediocre-stayer paradigm in Table 6 reveals that there are more stayers in S_2, and more movers in S_3. This inference is in line with the claim by Adomi and Mordi (2003) for S_2, but not S_3. A comparison of the staff categories indicates that the research performance of academics in S_3 outstrips those in S_1 and S_2, and that academics in S_1 did better than those in S_2 ($M_{12}+M_{13}>M_{22}+M_{23}$). Research publications in the university system are characterised by joint or collaborative authorship. Since this can happen, the possible explanation for the performance of S_1 over S_2 is possibly the research advice that those of S_1 benefit from members of S_3 by way of mentorship/supervision, which may culminate in co-authorship of research publication in reputable journals. The high proportion of stayers in S_2 does not portend a good omen for the system. This is because the task of mentorship and postgraduate supervision begins at S_2. Furthermore, the presence of stayers in S_3 does not reflect the desired expectation of academic accomplishment in terms of prestige and reputation ascribed to that category. It is expected that university as a centre of excellence, should ensure that mentorship and postgraduate supervision are provided by the best brains (or movers). Mentorship and postgraduate supervision are core duties of the staff in S_3. In at least four instances, g_i, $i=15,47,67,74$, an academic in S_3 is tagged a 'stayer'. This inference has a far-reaching implication on the horizontal mobility of academic staff in that category from one university to another and in the appointment of external assessors, external examiners and granting sabbatical. This is partially explained by considerations based on criteria other than publications (such as teaching/professional experience, conferences, administrative experience and general contribution) in the guidelines. Presumably, such academics may have earned more points from these other criteria to warrant their promotion to the professorial cadre. To improve on the quality of academics in this setting, there is a need to review the extant guidelines on appointments and promotions. In this sense the decision rule based on the calibrated LR model provides an early warning signal for the appropriate supervisory authorities to act.

Finally, this study demonstrates how to shortlist applicants for interview for a vacancy in the system as follows. Suppose that there are vacancies in the system and five applicants have submitted their CVs for consideration. Let the features extracted from their CV be the following: $\mathbf{x}^1=(2,0,1,12,14,9)$, $\mathbf{x}^2=(0,0,6,1,40,5)$, $\mathbf{x}^3=(0,0,9,1,51,5)$, $\mathbf{x}^4=(0,1,0,0,48,4)$, $\mathbf{x}^5=(0,1,3,2,22,8)$.

Table 7: Decision to Shortlist Applicants

Serial number	Class-conditional probability			Prediction	Decision to shortlist
1	0.1121	0.2566	0.6313	3	All categories
2	0.2408	0.3538	0.4055	3	All categories
3	0.1584	0.3069	0.5346	3	All categories
4	0.3966	0.3558	0.2476	1	S_1 only
5	0.2830	0.3630	0.3539	2	S_1 and S_2 only

Table 7 shows the categories for which the applicants can be shortlisted based on the calibrated LR model. Since a university is known by the quality of staff it attracts, the order of preference for the applicants' features is: $\mathbf{x}^1 \succ \mathbf{x}^3 \succ \mathbf{x}^2 \succ \mathbf{x}^5 \succ \mathbf{x}^4$, where the symbol $\succ$ means 'is preferred to'. This preference is determined by the category predictions and ties are differentiated by the individual posterior class-conditional probability. Although $\mathbf{x}^3$ has the highest number of publications (61 publications), yet it was not the most preferred. Rather $\mathbf{x}^1$ with just 29 publications is the most preferred applicant, because it has the highest number of articles in the SJR list with two articles in Q1 journals. Thus, the LR classifier is sensitive to journal quality.

5. Conclusion and Recommendations

This paper has made a giant stride in using the notion of the mover-mediocre-stayer subdivisions within the observable rank categories of academic staff to describe heterogeneity in the Nigerian university system. The study utilised the PCA method and the LR discriminative classifier. The PCA method has the potential to eliminate redundancy by reducing the size of the dataset. The study shows that the LR classifier, which is sensitive to journal quality (as ranked by Scimago), could be used as a tool to shortlist applicants for vacancies and can serve as a warning signal for supervisory authorities to improve on their guidelines for appointments and promotions. The obtained results reinforce the suitability of the LR method in discriminant analysis. It can be concluded that: regardless of the rank categories, the university system is made up of movers, mediocre and stayers.

One may be tempted to think that having a PhD is not a strong reason for research proficiency. This is because the category S_2, where PhD is a major requirement, has the highest percentage of stayers (66.66%), that is, less research performance. Nonetheless, this does not preclude the fact that an individual career path may be enhanced with possessing a doctorate degree. University Councils may decide to advance the standard for appointments and promotions, albeit with opposition from the trade union to which academics (in federal and state universities) in Nigeria belong. If there is a reason to believe that possessing a PhD and the type of university could affect research proficiency, then these variables could be included as factors in the LR model. Further, the quantity and quality of publications do not automatically translate to promotion, but the length of service experience may influence the position of an individual in the system. However, there were no data available to verify this claim. Further extension can take account of these variables to construct a non-parametric model for the problem. As earlier

mentioned, academics that their grade levels could not be ascertained from the RG and the Google Scholar platforms were not included in the training sample. Rather than constructing the LR classifier based on the labelled training sample, the study may be taken further by exploiting additional information contained in the research attributes of unlabelled sample via the development of a semi-supervised classification model.

Finally, the need to revisit the guidelines for appointments and promotions in Nigerian universities is recommended. The current practice where more emphasis is placed on publications in national journals with or without indexing should be discouraged. The allocation of equal point to published articles irrespective of the journal they are published should be reviewed. Publications in high Impact Factor (IF) journals should get a higher point than journals with low IF or without IF. There should be (financial) incentives for academics who publish in high IF journals. Finally, there is a need for the NUC to maintain a comprehensive database of the profile of academics in Nigerian universities. These recommendations will go a long way to improve research performance of Nigerian academics.

References

Adomi, E. E.and Mordi, C. (2003). Publication in foreign journals and promotion of academics in Nigeria, *Learned Publishing*, 16(4), 259–263. http://dx.doi.org/10.1087/095315103322421991.

Affes, Z. and Hentati-Kaffel, R. (2019). Predicting US banks bankruptcy: logit versus canonical discriminant analysis, *Computational Economics,* 54, 199–244. https://doi.org/10.1007/s10614-017-9698-0.

Arevalillo, J. M. and Navarro, H. (2012). A study of the effect of kurtosis on discriminant analysis under elliptical populations. *Journal of Multivariate Analysis*, 107, 53 – 63.

Bartholomew, D. J., Forbes, A. F. and McClean, S. I. (1991). *Statistical Techniques for Manpower Planning* (2nd ed.), John Wiley & Sons, Chichester.

Bi, Y. and Jeske, D. R. (2010). The efficiency of logistic regression compared to normal discriminant analysis under class-conditional classification noise, *Journal of Multivariate Analysis,* 101, 1622 – 1637.

Boudaghi, E. and Saen, R. F. (2014). Developing a model for determining optimal η in DEA-discriminant analysis for predicting suppliers' group membership in supply chain, *OPSEARCH*. DOI: 10.1007/s12597-014-0173-6.

Chauhan, S., Darvishzadeh, R., Boschetti, M. and Nelson, A. (2020). Discriminant analysis for lodging severity classification in wheat using RADARSAT-2 and Sentinel-1 data, *ISPRS Journal of Photogrammetry and Remote Sensing,* 164, 138 – 151.

Choubin, B., Moradi, E., Golshan, M., Adamowski, J., Sajedi-Hosseini, F. and Mosavi, A. (2019). An ensemble prediction of flood susceptibility using multivariate discriminant analysis, classification and regression trees, and support vector machines, *Science of the Total Environment,* 651, 2087 – 2096.

Cornell-Farrow, S. and Garrard, R. (2020). Machine learning classifiers do not improve the prediction of academic risk: Evidence from Australia, *Communications in Statistics:*

Case Studies, Data Analysis and Applications. DOI: 10.1080/2373748 4.2020.1752849.

Dattalo, P. (1995). A comparison of discriminant analysis and logistic regression, *Journal of Social Service Research* 19, (3-4), 121 – 144.

De Feyter, T. (2006). Modelling heterogeneity in manpower planning: dividing the personnel system into more homogeneous subgroups, *Applied Stochastic Models in Business and Industry,* 22, 321 – 334.

Deming, K., Yaoyao, C., Lingfu, K. and Shutao, W. (2018). Classification of oil pollutants based on excitation-emission matrix fluorescence spectroscopy and two-dimensional discriminant analysis, *Spectrochemical Acta Part A: Molecular and Biomolecular Spectroscopy*. https://doi.org/10.1016/j.saa.2019. 117799.

Demir, S.B. (2018). Predatory journals: Who publishes in them and why? *Journal of Informetrics,* 12, 1296–1311. DOI: https://doi.org/10.1016/j.joi.2018.10.008

Ekhosuehi, V.U. (2017). On graduation of enrolment size in a multi-echelon educational system, *Mathematica Applicanda [Matematyka Stosowana],* 45(2), 119 – 126. DOI: 10.14708/ma.v45i1.4378.

Ekhosuehi, V.U. and Omosigho, S.E. (2018). The use of certain staffing requirements as a means of benchmarking academic staff structure, *Mathematica Applicanda [Matematyka Stosowana],* 46(2), 259 - 272. DOI: 10.14708/mav46i2.5175.

Gao, J. and Li, L. (2019). A robust geometric mean-based subspace discriminant analysis feature extraction approach for image set classification, *Optik – International Journal for Light and Electron Optics,* 199, 163368.

Gaynanova, I. and Wang, T. (2019). Sparse quadratic classification rules via linear dimension reduction, *Journal of Multivariate Analysis,* 169, 278-299.

Hardle, W. andSimar, L. (2007). *Applied Multivariate Statistical Analysis* (2nded.), Springer-Verlag, Berlin.

Harris, G. and Kaine, G. (1994). The determinant of research performance: a study of Australian university economists, *Higher Education,* 27, 191 – 201.

Katos, V. (2007). Network intrusion detection: evaluating cluster, discriminant and logit analysis, *Information Sciences,* 177, 3060 – 3073.DOI: 10.1016/j.ins.2007.02. 034

Manjunath, B.G., Frick, M. and Reiss, R.-D. (2012). Some notes on extremal discriminant analysis, *Journal of Multivariate Analysis,* 103, 107 – 115.

Maugis, C., Celeux, G. and Martin-Magniette, M. L. (2011). Variable selection in model-based discriminant analysis, *Journal of Multivariate Analysis,* 102, 1374 – 1387.

McCullagh, P. (1980). Regression models for ordinal data, *Journal of the Royal Statistical Society*, Series B, 42 (2), 109-142.

Minka, T.P. (2004). A comparison of numerical optimizers for logistic regression. *Technical Report* 758, Carnegie Mellon University.

Nwosu, D.F., Onyeagu, S.I., Mbegbu, J.I. and Ekhosuehi, V.U. (2016). On refined principal component method for factor analysis, *Journal of the Nigerian Statistical Association,* 28, 16 – 29.

Nwosu, D.F., Ekhosuehi, V.U. and Mbegbu, J.I. (2020). Performance of some factor analysis techniques, *Annals of Data Science,* 7, 209-242. DOI 10.1007/s40745-020-00260-6.

Oravec, M., Beganovic, A., Gal, L., Ceppan, M. and Huck, C.W. (2019). Forensic classification of black inkjet prints using Fourier transform near-infrared spectroscopy and linear discriminant analysis, *Forensic Science International,* 299, 128 – 134.

Parmet, Y., Schechtman, E. and Sherman, M. (2010). Factor analysis revisited – How many factors are there? *Communications in Statistics – Simulation and Computation,* 39**,** 1893 – 1908.

Press, S.J. and Wilson, S. (1978). Choosing between logistic regression and discriminant analysis, *Journal of the American Statistical Association,* 73, 699 – 705.

Rencher, A.C. (2002). *Method of Multivariate Analysis (2nd ed.),* John Wiley and Sons Inc., New York.

Renjith, G., Mary, D.P., Soe, K., Wan, M.Y., Beh, H.C., Phuah, W.H. andTseu, A.T.H. (2019). Sex estimation by discriminant function analysis using anatomical location of mental foramen, *Forensic Science International: Reports*, 1, 100018.

Setser, A.L. and Smith, R.W. (2018). Comparison of variable selection methods prior to linear discriminant analysis classification of synthetic phenethylamines and tryptamines, *Forensic Chemistry*. https://doi.org/10.1016/j.forc.2018.10.002

Tavernier, J., Simm, J., Meerbergen, K., Wegner, J.K., Ceulemans, H. and Moreau, Y. (2019). Fast semi-supervised discriminant analysis for binary classification of large data sets, *Pattern Recognition,* 91, 86 – 99.

Udom, A.U. and Ebedoro, U.G. (2019). On multinomial hidden Markov model for hierarchical manpower systems, *Communications in Statistics – Theory and Methods*. DOI: 10.1080/ 03610926.2019.1650185.

Xia, C., Yang, S., Huang, M., Zhu, Q., Guo, Y. and Qin, J. (2019). Maize seed classification using hyperspectral image coupled with multi-linear discriminant analysis. *Infrared Physics & Technology*. DOI: https://doi.org/10.1016/j.infrared. 2019.103077.

You, X., Xu, J., Yuan, W., Jing, X.-Y., Tao, D. and Zhang, T. (2019). Multi-view common component discriminant analysis for cross-view classification, *Pattern Recognition,* 92, 37 – 51.

Zheng, J., Wang, H., Song, Z. and Ge, Z. (2019). Ensemble semi-supervised Fisher discriminant analysis model for fault classification in industrial processes, *ISA Transactions,* 92, 109 – 117.

Zhou, F., Guo, H.-C., Ho, Y.-S.and Wu, C.-Z. (2007). Scientometric analysis of geostatistics using multivariate methods, *Scientometrics*, 73(3), 265 – 279. DOI: 10.1007/s11192-007-1798-5.

Zou, M., Sun, C., Liang, S., Sun, Y., Li, D., Li, L., Fan, L., Wu, L.and Xia, W. (2018). Fisher discriminant analysis for classification of autism spectrum disorders based on folate-related metabolism markers. *Journal of Nutritional Biochemistry*. DOI: 10.1016/j.jnutbio.2018.09.023.

Appendix: Training data set of selected academics from different universities

Individual index	Research features						Staff category	RG score
	Q1	Q2	Q3	Q4	Others	h-index		
1	0	0	0	0	12	3	S2	0
2	3	2	4	3	99	8	S3	18.41
3	0	0	1	1	16	3	S1	4.18
4	0	1	1	2	34	7	S3	10.2
5	0	0	0	0	7	1	S2	3.36
6	4	1	5	5	31	8	S3	13.76
7	1	1	1	1	11	3	S3	6.76
8	0	1	0	0	7	2	S1	2.34
9	0	0	0	2	5	2	S2	3.15
10	0	0	3	5	9	10	S2	5.88
11	2	1	8	12	30	9	S3	14.4
12	0	1	0	0	6	2	S1	1.74
13	0	0	0	1	33	5	S2	7.73
14	6	3	4	3	18	12	S3	18.62
15	1	0	0	0	14	4	S3	3.24
16	33	35	37	8	68	25	S3	34.28
17	0	2	2	3	5	2	S1	7.18
18	1	3	1	0	8	4	S1	5.54
19	2	1	0	0	22	5	S2	7.9
20	1	5	5	2	33	7	S2	15.56
21	1	0	0	1	4	1	S1	2.58
22	0	0	0	0	27	4	S1	2.78
23	8	5	11	0	16	11	S3	20.3
24	0	0	2	0	3	2	S1	3.6
25	9	7	4	7	20	10	S3	20.31
26	0	0	2	0	33	5	S1	4.58
27	0	0	0	0	10	1	S1	1.83
28	0	3	8	9	20	10	S3	13.56
29	0	0	1	0	12	2	S1	2.47
30	0	0	0	1	37	1	S1	6.21
31	0	0	0	1	29	4	S2	6.67
32	3	0	1	1	9	5	S1	12.16
33	0	1	2	4	22	9	S3	8.66
34	2	0	0	2	15	3	S3	7.92
35	1	9	4	1	55	10	S3	12.13
36	0	2	4	6	19	12	S1	8.84
37	0	3	6	0	10	7	S2	7.91
38	5	13	9	7	59	13	S1	21.08
39	0	7	7	5	24	14	S3	11.15
40	6	9	17	1	24	15	S3	24.21
41	2	0	1	0	29	6	S2	9.0
42	0	0	0	0	12	6	S2	3.73
43	0	0	1	0	5	13	S1	2.37

44	3	0	2	0	3	4	S1	9.69
45	0	6	1	0	10	7	S1	13.31
46	3	10	12	1	23	9	S2	20.64
47	0	1	0	0	34	5	S3	6.04
48	14	4	10	1	16	20	S3	28.22
49	0	0	0	1	7	0	S1	1.71
50	3	0	0	0	7	0	S1	8.52
51	0	0	0	0	20	3	S1	3.82
52	3	4	3	0	55	6	S1	9.69
53	0	1	1	1	9	4	S2	8.09
54	0	0	0	2	24	7	S1	2.73
55	1	0	1	0	5	2	S1	3.44
56	2	0	7	2	2	4	S1	12.45
57	4	4	5	1	33	12	S2	17.94
58	14	4	7	0	22	17	S3	23.13
59	3	6	9	0	5	9	S3	17.87
60	0	0	2	0	36	2	S2	7.4
61	0	1	1	0	26	5	S2	7.37
62	3	1	3	1	5	3	S3	10.71
63	0	0	3	0	29	11	S3	5.66
64	0	0	0	0	7	3	S2	1.21
65	0	0	0	1	3	2	S1	1.71
66	0	5	2	2	36	13	S3	14.02
67	0	2	0	0	10	3	S3	2.17
68	0	0	0	0	22	2	S2	5.45
69	0	2	1	3	14	7	S2	8.18
70	0	0	0	0	5	8	S1	1.98
71	5	4	10	5	24	11	S3	19.09
72	1	0	0	5	24	4	S2	7.67
73	1	3	0	0	19	7	S2	3.99
74	1	0	0	0	11	3	S3	5.8
75	4	13	7	9	25	10	S3	23.16