CHAPTER FIVE

PRINCIPAL COMPONENT ANALYSIS AND ZERO-HEAVY DATA

5.1 Introduction

Previous chapters of this dissertation have focused on the modeling and analysis of zero-heavy data in the regression framework. There are some interesting questions in environmental studies in which the growth or reproduction rates of multi-species are measured at numerous sites. This also often gives rise to multivariate data and involves multivariate methods.

Questions of interests are:

- How is information in the correlation structure between species affected by zeroes?
- Can multivariate methods be adopted to handle the problem of zero heavy data?

Since the correlation structure of the data is of major interest in multivariate data analysis, this chapter will concentrate on methods that explore the correlation structure. Principal component analysis (PCA) and factor analysis (FA) are two of the most common methods used to investigate the correlation structure. PCA decomposes the correlation (or variance-covariance) into eigenvalues and orthogonal eigenvectors which explain the variation and interrelationship of the variables in the data. Factor analysis is another method that also explores the correlation structure of the data. It is a model driven analysis where the estimated factor may have “meaningful” interpretations.

The purpose of this Chapter is to do an exploratory study on the effect of zeroes on correlation structure. In the following sections, the backgrounds of principal component analysis and factor analysis are discussed.

5.2 Principal Component Analysis and Factor Analysis

The main idea of principal component analysis is to reduce the dimensionality of a data set that may contain many highly correlated variables, while retaining as much as possible of the variations in the data set. This also makes the interpretation of the data easier. Algebraically, principal components are computed using either the correlation matrix ρ or the covariance matrix Σ which will generate an eigenvectors e_i and eigenvalues λ_i. The covariance matrix can be expressed as follows:

$$\Sigma = \sum_{i=1}^{p} \lambda_i e_i e_i'$$

The eigenvalues represent the proportion of the total variation explained which will provide information about the correlation of the variables.

Factor analysis is another model that explores the variance-covariance matrix. Let Y ($p \times 1$) be a vector of observable random variables, then the factor analysis model is:

$$Y = LF + \varepsilon,$$

where L is the factor loading matrix ($p \times m$), F is a vector ($m \times 1$) of latent variables called the factors, and ε is a ($p \times 1$) vector of errors. In the case of an orthogonal factor model, the following assumptions are made:

- F and ε are independent;
- E(F) = 0, Cov(F) =I, and

- $E(\varepsilon) = 0$, $Cov(\varepsilon) = \Psi$, where Ψ is a diagonal matrix.

An orthogonal factor model is similar to a regression model except the regressors are unknown (Longford, 1993). Here again, the analysis involves the decomposition of Σ, the variance-covariance matrix of Y. The covariance structure is as follows:

- $Cov(Y) = \Sigma = LL' + \Psi$, and
- $Cov(Y,F) = L$

This implies the followings:

- $\text{var}(Y_i) = l_{i1}^2 + l_{i2}^2 + \cdots + l_{im}^2 + \psi_i$;
- $\text{cov}(Y_i, Y_k) = l_{i1}l_{k1} + l_{i2}l_{k2} + \cdots + l_{im}l_{km}$, and
- $\text{cov}(Y_i, F_j) = l_{ij}$

where l_{ij} is the ij^{th} element of the loading matrix L, and ψ_i is the diagonal element of matrix Ψ. The loadings and factors can be obtained using different estimation techniques, but different techniques will give different solutions. For example, maximum likelihood, regression, or even principal component analysis can serve as an estimation method for obtaining the factor and its loadings in the factor analysis. In fact, the models of principal component analysis and factor analysis share some similarities, and principal component analysis can be viewed as a special case of the factor analysis model (Basilevsky, 1994). A factor analysis model with $\Psi = 0$ and loading $L = \left[e_i\sqrt{\lambda_i} \quad \cdots \quad e_p\sqrt{\lambda_p} \right]$ is exactly the principal component with eigenvalues λ_i and eigenvectors e_i $(i = 1,\ldots, p)$. It is clear now that any changes in the correlation structure will alter the size of the eigenvalues and loadings of these analyses, and hence, affect the results of these analyses. Both the variance-covariance matrix and the correlation matrix can be used in these two analyses.

However, to avoid any confusion, only the correlation matrix will be used in the following discussion.

5.3 Principal Component Analysis for Zero-Heavy Data

As mentioned in previous chapters, zero-heavy data is very common in environmental studies or toxicity testing. In many cases, it is clear that the zeroes are created by a mechanism other than the one that creates the non-zero data. For example, most zeroes are due to mortality, and fertility create non-zero data. The mixture model idea can then be applied here by letting the observation Y be a product X*Z, where X can be any non-negative variable measuring growth or fecundity, and Z is a binary variable indicating mortality. If the animals survive, then Z = 1; and if the animals die, then Z = 0. In most chronic effect studies, the interest is mainly focused on either growth or fecundity, i.e. the complete X variable, and researchers are more interested in the correlation structure of X than those of the Z variables. In this case, X is actually missing when the animals die (Z = 0) and Y is actually the observed part of X combined with Z = 0.

In the following section, a simulation study is carried out to examine the effect of zeroes on principal component analysis. This is done by comparing the principal component analysis of Y and X. Since the complete X is not observable, the correlation structure has to be recovered by an estimation method. In order to recover the correlation structure of X, two different methods of replacing the zeroes are used. They are the mean substitution method and the EM algorithm (Little and Rubin, 1987). The mean method simply substitutes the zeroes with the means of the variables, and the EM algorithm iteratively updates the estimation until it converges.

5.4 Simulation Studies

In order to generate the mortality mechanism, random binary numbers are generated based on a fixed mortality probability. However, in real life situation, the mortality mechanism may be related to the mechanism which creates the non-zero data, for instance, the probability of mortality may increase as the growth increases. This may happen if the toxicants affect the mortality and fecundity biological mechanisms in the same way. Therefore, a second set of simulation is done by relating the mortality probability to the mean of the observed variables.

5.4.1 Simulation I (Constant Probability of Mortality)

The first set of simulations is carried out as a factorial design with 3 factors and 2 levels for each factor. The three factors are the number of variables (p), the probability of mortality (PM), and the noise level in the correlation matrix (N). In order to make sure the first and second eigenvalues are well separated, a block diagonal structure is used for the true correlation matrix. The predetermined correlation matrices are shown in Figure 5.1-5.4. The noise level is defined as the relative size of the correlation of the off diagonal block elements. The notation used in the first set of simulation are listed in Table 5.1.

Table 5.1. Explanation of the notations for simulation I

n	Number of observations: 40 and 100
P	Number of variables 2 levels: p=5, p=10
PM	PM: Probability of Mortality i.e., Prob(0) 3 levels:PM=0, PM=0.2, PM=0.5
True	From the true correlation
With 0	From the data with zeroes (Y)
Mean	X re-constructed by Mean substitute method
EM	X re-constructed by EM algorithm (iterative procedure)
High Noise	Relatively higher correlation at the off diagonal block
Low Noise	Relatively lower correlation at the off diagonal block

$$\begin{bmatrix} 1.0 & 0.89 & 0.9 & 0.9 & 0.9 & 0.9 & 0.01 & 0.01 & 0.01 & 0.01 \\ & 1.0 & 0.95 & 0.9 & 0.9 & 0.9 & 0.01 & 0.02 & 0.03 & 0.01 \\ & & 1.0 & 0.85 & 0.9 & 0.96 & 0.01 & 0.01 & 0.01 & 0.01 \\ & & & 1.0 & 0.8 & 0.9 & 0.01 & 0.01 & 0.04 & 0.01 \\ & & & & 1.0 & 0.9 & 0.01 & 0.01 & 0.01 & 0.01 \\ & & & & & 1.0 & 0.01 & 0.01 & 0.01 & 0.01 \\ & & & & & & 1.0 & 0.95 & 0.89 & 0.9 \\ & & & & & & & 1.0 & 0.89 & 0.9 \\ & & & & & & & & 1.0 & 0.7 \\ & & & & & & & & & 1.0 \end{bmatrix}$$

Figure 5.1. 10x10 Correlation matrix (symmetric) with block diagonal elements
Relatively lower correlation at off diagonal elements

$$\begin{bmatrix} 1.0 & 0.89 & 0.9 & 0.9 & 0.9 & 0.9 & 0.4 & 0.4 & 0.4 & 0.4 \\ & 1.0 & 0.95 & 0.9 & 0.9 & 0.9 & 0.4 & 0.4 & 0.4 & 0.4 \\ & & 1.0 & 0.85 & 0.9 & 0.96 & 0.4 & 0.4 & 0.4 & 0.4 \\ & & & 1.0 & 0.8 & 0.9 & 0.4 & 0.4 & 0.4 & 0.4 \\ & & & & 1.0 & 0.9 & 0.4 & 0.4 & 0.4 & 0.4 \\ & & & & & 1.0 & 0.4 & 0.4 & 0.4 & 0.4 \\ & & & & & & 1.0 & 0.95 & 0.89 & 0.9 \\ & & & & & & & 1.0 & 0.89 & 0.9 \\ & & & & & & & & 1.0 & 0.7 \\ & & & & & & & & & 1.0 \end{bmatrix}$$

Figure 5.2. 10x10 Correlation matrix (symmetric) with block diagonal elements
Relatively higher correlation at off diagonal elements

$$\begin{bmatrix} 1.0 & 0.9 & 0.8 & 0.04 & 0.09 \\ & 1.0 & 0.7 & 0.02 & 0.01 \\ & & 1.0 & 0.009 & 0.05 \\ & & & 1.0 & 0.08 \\ & & & & 1.0 \end{bmatrix}$$

Figure 5.3. 5×5 Correlation (symmetric) matrix with block diagonal elements
Relatively lower correlation at off diagonal elements

$$\begin{bmatrix} 1.0 & 0.9 & 0.8 & 0.3 & 0.4 \\ & 1.0 & 0.7 & 0.2 & 0.1 \\ & & 1.0 & 0.009 & 0.05 \\ & & & 1.0 & 0.08 \\ & & & & 1.0 \end{bmatrix}$$

Figure5.4. 5×5 Correlation matrix (symmetric) with block diagonal elements
Relatively higher correlation at off diagonal elements

The multivariate normal data are generated using a SAS program in which random binary data are also generated to simulate the mortality effect, which is assumed to be random with probability PM. The different combinations of runs are coded in Table 5.2. and Table 5.3. The observed data Y is generated by an element-wise multiplication of the two data matrix X and Z, i.e. Y=X#Z. Principal component analysis is then performed on Y and X. A SAS macro for the simulation procedure is presented in Appendix D.

Table 5.2. The three factors in the simulation study

	High (+1)	Low (-1)
P	10	5
Noise	High	Low
PM	0.5	0.2

Table 5.3. The combinations of the different factors in the study

Run number.	p	Noise	PM
1	1	1	1
2	1	1	-1
3	1	-1	1
4	1	-1	-1
5	-1	1	1
6	-1	1	-1
7	-1	-1	1
8	-1	-1	-1

5.4.2 Results for simulation I (Constant Mortality Probability)

Since the eigenvalues indicate of how much variation is explained, they are the focus of the analysis. The eigenvalues of the principal component analyses of the generated data are presented in Table 5.4 – Table 5.7. The results of each run are the average of 100 values. The percentage of variation explained by the first two eigenvalues are also listed in Table 5.8 and Table 5.9.

Table 5.4. Eigenvalues for the simulated data (means of 100) in simulation I
n=100, p=5; S.E. in ()

p	Eigenvalues (ordered)			1	2	3	4	5
5	Low Noise	**True**		**2.613**	**1.791**	**0.318**	**0.201**	**0.077**
		PM=0		2.671 (0.0860)	1.733 (0.0756)	0.323 (0.0563)	0.200 (0.0286)	0.073 (0.0140)
		PM=0.2	With 0	1.366 (0.1027)	1.140 (0.0638)	0.969 (0.0531)	0.835 (0.0591)	0.690 (0.0650)
			Mean	2.611 (0.2109)	1.300 (0.1560)	0.492 (0.0776)	0.350 (0.0559)	0.248 (0.0525)
			EM	2.676 (0.0939)	1.727 (0.0851)	0.330 (0.0667)	0.198 (0.0378)	0.070 (0.0163)
		PM=0.5	With 0	1.315 (0.0742)	1.133 (0.0537)	0.996 (0.0426)	0.862 (0.0502)	0.694 (0.0716)
			Mean	2.824 (0.1602)	0.748 (0.0680)	0.607 (0.0634)	0.484 (0.0607)	0.337 (0.0567)
			EM	2.753 (0.1674)	1.693 (0.1773)	0.372 (0.1316)	0.149 (0.0663)	0.033 (0.0242)
	High Noise	**True**		**2.805**	**1.657**	**0.329**	**0.204**	**0.005**
		PM=0		2.811 (0.1546)	1.641 (0.1418)	0.340 (0.0519)	0.202 (0.0342)	0.005 (0.0009)
		PM=0.2	With 0	1.600 (0.1296)	1.173 (0.0878)	0.948 (0.0665)	0.770 (0.0857)	0.509 (0.0833)
			Mean	2.226 (0.1725)	1.075 (0.1162)	0.735 (0.0936)	0.571 (0.0709)	0.394 (0.0635)
			EM	2.811 (0.1556)	1.645 (0.1446)	0.342 (0.0568)	0.197 (0.0368)	0.005 (0.0015)
		PM=0.5	With 0	1.334 (0.0786)	1.143 (0.0668)	0.985 (0.0512)	0.848 (0.0528)	0.690 (0.0729)
			Mean	2.698 (0.2108)	0.767 (0.0851)	0.636 (0.0681)	0.527 (0.0660)	0.372 (0.0675)
			EM	2.839 (0.2756)	1.634 (0.2382)	0.384 (0.1694)	0.127 (0.0765)	0.015 (0.0118)

Table 5.5. First 5 eigenvalues for the simulated data (means of 100) in simulation I
n =100, p=10; S.E. in ()

p	Eigenvalues (ordered)			1	2	3	4	5
10	Low Noise	**True**		**5.487**	**3.618**	**0.304**	**0.207**	**0.135**
		PM=0		5.590 (0.1847)	3.517 (0.1744)	0.317 (0.4444)	0.206 (0.0381)	0.135 (0.0231)
		PM=0.2	With 0	1.770 (0.1534)	1.443 (0.0867)	1.239 (0.0691)	1.098 (0.0583)	0.985 (0.0448)
			Mean	4.682 (0.3979)	2.481 (0.2800)	0.662 (0.0904)	0.528 (0.0618)	0.427 (0.0542)
			EM	5.591 (0.1865)	3.522 (0.1750)	0.316 (0.0482)	0.207 (0.0356)	0.137 (0.0248)
		PM=0.5	With 0	1.559 (0.0939)	1.364 (0.0612)	1.235 (0.0589)	1.127 (0.0472)	1.021 (0.0462)
			Mean	4.693 (0.3757)	1.099 (0.1247)	0.861 (0.0832)	0.735 (0.0733)	0.636 (0.0632)
			EM	5.594 (0.3042)	3.456 (0.3100)	0.390 (0.2082)	0.231 (0.0867)	0.148 (0.0611)
	High Noise	**True**		**6.724**	**2.385**	**0.300**	**0.209**	**0.131**
		PM=0		6.785 (0.3843)	2.333 (0.3225)	0.309 (0.0501)	0.207 (0.0374)	0.132 (0.0233)
		PM=0.2	With 0	1.801 (0.1865)	1.389 (0.0861)	1.221 (0.0657)	1.099 (0.0531)	0.990 (0.0446)
			Mean	6.129 (0.4281)	1.488 (0.2384)	0.579 (0.0880)	0.444 (0.0680)	0.354 (0.0522)
			EM	6.798 (0.3914)	2.329 (0.3332)	0.305 (0.0534)	0.203 (0.0354)	0.133 (0.0381)
		PM=0.5	With 0	1.566 (0.0985)	1.372 (0.0639)	1.233 (0.0572)	1.127 (0.0474)	1.017 (0.0440)
			Mean	5.163 (0.4052)	0.935 (0.1122)	0.780 (00774)	0.680 (0.0680)	0.588 (0.0640)
			EM	6.751 (0.4248)	2.322 (0.3540)	0.324 (0.0847)	0.204 (0.0504)	0.139 (0.0325)

Table 5.6. Eigenvalues for the simulated data (means of 100) in simulation I
n =40, p=5; S.E. in ()

p	Eigenvalues (ordered)			1	2	3	4	5
5	Low Noise	**True**		**2.613**	**1.791**	**0.318**	**0.201**	**0.077**
		PM=0		2.757 (0.1997)	1.655 (0.1658)	0.327 (0.0855)	0.190 (0.0468)	0.070 (0.0232)
		PM=0.2	With 0	1.530 (0.1603)	1.201 (0.0932)	0.958 (0.0812)	0.763 (0.0932)	0.548 (0.1062)
			Mean	2.647 (0.3308)	1.291 (0.2391)	0.538 (0.1224)	0.325 (0.0751)	0.198 (0.0526)
			EM	2.785 (0.2111)	1.633 (0.1845)	0.343 (0.1060)	0.179 (0.0548)	0.060 (0.0244)
		PM=0.5	With 0	1.482 (0.1363)	1.203 (0.0854)	0.968 (0.0724)	0.777 (0.0880)	0.570 (0.0854)
			Mean	2.848 (0.2687)	0.803 (0.1101)	0.607 (0.0968)	0.452 (0.0891)	0.289 (0.0823)
			EM	2.775 (0.3579)	1.579 (0.2923)	0.474 (0.2641)	0.127 (0.0773)	0.045 (0.0404)
	High Noise	**True**		**2.805**	**1.657**	**0.329**	**0.204**	**0.005**
		PM=0		2.825 (0.2284)	1.629 (0.2197)	0.356 (0.0863)	0.185 (0.0475)	0.005 (0.0014)
		PM=0.2	With 0	1.716 (0.1823)	1.238 (0.1134)	0.936 (0.0842)	0.688 (0.0942)	0.423 (0.0839)
			Mean	2.229 (0.2699)	1.140 (0.1742)	0.765 (0.1199)	0.539 (0.0939)	0.327 (0.0965)
			EM	2.831 (0.2303)	1.634 (0.2278)	0.356 (0.0892)	0.170 (0.0496)	0.010 (0.0052)
		PM=0.5	With 0	1.497 (0.1269)	1.209 (0.0905)	0.973 (0.0760)	0.775 (0.0797)	0.546 (0.0967)
			Mean	2.690 (0.3248)	0.870 (0.1421)	0.641 (0.1043)	0.475 (0.0906)	0.324 (0.0870)
			EM	2.897 (0.3739)	1.508 (0.3082)	0.451 (0.2364)	0.117 (0.0854)	0.028 (0.0256)

Table 5.7. First 5 eigenvalues for the simulated data (means of 100) in simulation I
n = 40, p = 10; S.E. in ()

p	Eigenvalues (ordered)			1	2	3	4	5
10	Low Noise	**True**		**5.487**	**3.618**	**0.304**	**0.207**	**0.135**
		PM=0		5.700 (0.2856)	3.404 (0.2877)	0.353 (0.0755)	0.204 (0.0441)	0.131 (0.0315)
		PM=0.2	With 0	2.091 (0.2190)	1.664 (0.1302)	1.383 (0.0976)	1.152 (0.0799)	0.975 (0.0775)
			Mean	4.660 (0.5531)	2.537 (0.4405)	0.796 (0.1551)	0.566 (0.0894)	0.437 (0.0830)
			EM	5.702 (0.2879)	3.413 (0.2896)	0.358 (0.0801)	0.206 (0.0472)	0.130 (0.0325)
		PM=0.5	With 0	1.918 (0.1635)	1.596 (0.0963)	1.361 (0.0839)	1.166 (0.0694)	1.000 (0.0726)
			Mean	4.626 (0.5501)	1.286 (0.1885)	0.989 (0.1414)	0.786 (0.0985)	0.646 (0.0856)
			EM	5.602 (0.4943)	3.192 (0.4474)	0.598 (0.2837)	0.299 (0.1183)	0.166 (0.0832)
	High Noise	**True**		**6.724**	**2.385**	**0.300**	**0.209**	**0.131**
		PM=0		6.827 (0.5066)	2.332 (0.4305)	0.314 (0.0721)	0.201 (0.0509)	0.125 (0.0343)
		PM=0.2	With 0	2.102 (0.2531)	1.627 (0.1368)	1.360 (0.1026)	1.153 (0.0769)	0.984 (0.0758)
			Mean	6.113 (0.6347)	1.574 (0.3587)	0.666 (0.1532)	0.476 (0.1007)	0.354 (0.0763)
			EM	6.833 (0.5216)	2.338 (0.4516)	0.319 (0.0798)	0.203 (0.0507)	0.126 (0.0364)
		PM=0.5	With 0	1.942 (0.1708)	1.596 (0.0978)	1.356 (0.0805)	1.171 (0.0747)	1.013 (0.0604)
			Mean	5.219 (0.6693)	1.144 (0.1911)	0.893 (0.1357)	0.717 (0.1100)	0.583 (0.0846)
			EM	6.555 (0.7427)	2.325 (0.6113)	0.526 (0.2944)	0.297 (0.1803)	0.159 (0.1033)

Table 5.8. The percentage of variation explained by the first two eigenvalues in the simulation study I, n =100

	EM (re-constructed X)		Mean (re-constructed X)		With 0 (Y)		No mortality (complete X)	
Run Number	first eigen-value	second eigen-value	first eigen-value	second eigen-value	first eigen-value	second eigen-value	first eigen-value	second eigen-value
1	67.5%	23.2%	51.6%	9.4%	15.7%	13.7%	67.9%	23.3%
2	67.9%	23.3%	61.3%	14.9%	18.0%	13.9%	67.9%	23.3%
3	55.9%	34.6%	46.9%	11.0%	15.6%	13.6%	55.9%	35.2%
4	55.9%	35.2%	46.8%	24.8%	17.7%	14.4%	55.9%	35.2%
5	56.8%	32.7%	54%	15.3%	26.7%	22.9%	56.2%	32.8%
6	56.2%	32.9%	44.5%	21.5%	32%	23.5%	56.2%	32.8%
7	55.1%	33.9%	56.5%	15%	26.3%	22.7%	53.4%	34.7%
8	53.5%	34.5%	52.2%	26%	27.3%	22.8%	53.4%	34.7%

Table 5.9. The percentage of variation explained by the first two eigenvalues in the simulation study I, n =40

	EM (re-constructed X)		Mean (re-constructed X)		With 0 (Y)		No mortality (complete X)	
Run Number	first eigen-value	second eigen-value	first eigen-value	second eigen-value	first eigen-value	second eigen-value	first eigen-value	second eigen-value
1	65.6%	23.2%	52.2%	11.4%	19.4%	15.0%	68.2%	23.3%
2	68.3%	23.4%	61.1%	15.7%	21.0%	16.3%	68.2%	23.3%
3	56.0%	31.9%	46.3%	12.9%	19.2%	16.0%	57.0%	34.0%
4	57.0%	34.1%	46.6%	25.4%	20.9%	16.6%	57.0%	34.0%
5	57.9%	30.2%	53.8%	17.4%	29.9%	24.2%	56.5%	32.6%
6	56.6%	32.7%	44.6%	22.8%	34.3%	24.8%	56.5%	32.6%
7	55.5%	31.6%	57%	16.1%	29.6%	24.1%	55.1%	33.1%
8	55.7%	32.7%	52.9%	25.8%	30.6%	24%	55.1%	33.1%

Only the percentages of variation explained by the first and second eigenvalues are shown in Table 5.8 and Table 5.9 because they account for about 90% of the variation of the complete X variable. To allow for comparison, the eigenvalues of the

re-constructed data and of the original data are shown at the right-hand side of Table 5.8 and Table 5.9. It is obvious from the "With Zero" columns of Tables 5.4-5.7 that the presence of zeroes highly influences the correlation structure, and hence, alters the results of the principal component analysis. The performance of the EM method is good because:

- the results generated from this method are very consistent;
- the method is also able to recover the correlation structure of the variable X, and
- when comparing the results of the EM algorithm to the no mortality case, the difference in their eigenvalues ranges from 3% to 0%.

On the other hand, the mean substitution method does not perform as well as the EM method. In particular, the values of the second eigenvalues are quite different for the no mortality and the mean case. When considering the first eigenvalues, the mean method does perform slightly better when p=5 than when p=10. This may be due to the relatively smaller number of data point in the case for p=5. So mean substitution method does not perform as well as the EM method.

5.4.3 Simulation II (Probability of Mortality related to the Mean of X)

In reality, the probability of mortality is often related to the observed variables. In order to further investigate the effect of these zero creating mechanisms, a second set of simulations is carried out by relating the mortality probability to the mean of the observed variables. An example for how the mortality probability is related to the mean of the observed variables is when growth or size of the species is the main research interest. Older organisms usually have a larger size, but older organisms also tend to have a higher mortality. So mortality is directly proportional to the size of the organisms as well as the age, in other words, mortality increases as the size and/ the age of the organism increase(s). A similar example can also be found in the case that involves reproduction, aging and mortality. Generally, the reproduction rate decreases as one ages,

and mortality also increases as one ages. Thus, a spurious relationship may develop indicating that mortality is inversely proportional to rate of reproduction. The second set of simulations uses mostly the same notation as the first simulation except the probability of mortality (PM) is designed to be either directly proportional or inversely proportional to the mean of the variable X to mimic the situations described above. The coding for the simulation is the same as Table 5.2 and Table 5.3 except for the probability of mortality (PM). "+1" is assigned to the case when PM is directly proportional to the mean of the non-zero data, and "–1" is assigned to the case when PM is inversely proportional to the mean of the non-zero data. The notations used in the second set of simulation are listed in Table 5.10. In the case of p = 10, the values of PM range from 0.15 to 0.5 and the mean of X ranges from 4 -14. In the case of p=5, the values of PM range from 0.2 to 0.5, the mean of X ranges from 4 – 10. The simulation is again done by a SAS program.

Table 5.10. Explanation of the notations for simulation II

n	Number of observations 40 and 100
P	Number of variables 2 levels p=5, p=10
PM	PM: Probability of Mortality i.e., Prob(0) 3 levels PM=0 PM∝ mean (1) PM∝1/mean(-1) (∝ represents proportional to)
True	From the true correlation
With 0	From the data with zeroes (Y)
Mean	X re-constructed by Mean substitute method
EM	X re-constructed by EM algorithm (iterative procedure)
High Noise	Relatively higher correlation at the off diagonal block
Low Noise	Relatively lower correlation at the off diagonal block

5.4.4 Results of Simulation II

The study again focuses on the eigenvalues obtained by the principal component analysis. The values of first five eigenvalues are presented in Table 5.11-Table 5.14, and the percentages of variation explained by the first two eigenvalues are presented in Tables 5.15-5.16. Again, the correlation structures of the observed data (Y) are highly affected by the zeroes. The percentage explained by the first two eigenvalues drops from over 90% to as low as 30% (run#2, n=100).

The EM algorithm once again out-performed the mean substitution method. Results obtained from the EM algorithm are very consistent. The values obtained from the EM method are very closed to the no mortality case. The different values of the mortality probability do not affect the results obtained from the EM algorithm. On the other hand, the mean method can only retain about 50% of variation by its first two eigenvalues in contrast to the 90% of the no mortality case. The mean method shows slightly different results in the two different of mortality probability (PM) cases. The first eigenvalues of the mean method explained less variation when the mortality probability is directly proportional to the mean of the observed variables (run #2,4,6,8) than when the mortality probability is inversely proportional to the mean of the observed variables (run#1,3,5,7).

Table 5.11. Eigenvalues for the simulated data (means of 100) in simulation II
n=100, p=5; S.E. in ()

p	Eigenvalues (ordered)			1	2	3	4	5
5	Low Noise	**True**		**2.613**	**1.791**	**0.318**	**0.201**	**0.077**
		PM=0		2.648 (0.0860)	1.765 (0.0680)	0.323 (0.0567)	0.190 (0.0306)	0.074 (0.0144)
		PM ∝1/Mean	With 0	1.309 (0.0856)	1.124 (0.0523)	0.983 (0.0415)	0.857 (0.0511)	0.727 (0.0618)
			Mean	1.520 (0.1202)	1.251 (0.1022)	0.924 (0.0779)	0.736 (0.0830)	0.570 (0.0767)
			EM	2.650 (0.1054)	1.760 (0.0838)	0.334 (0.0681)	0.182 (0.0387)	0.073 (0.0219)
		PM∝Mean	With 0	1.317 (0.0872)	1.120 (0.0549)	0.988 (0.0480)	0.859 (0.0544)	0.716 (0.0666)
			Mean	1.815 (0.1743)	1.129 (0.0865)	0.920 (0.0647)	0.712 (0.0957)	0.424 (0.1013)
			EM	2.661 (0.1033)	1.757 (0.0819)	0.334 (0.0720)	0.179 (0.0437)	0.069 (0.0197)
	High Noise	**True**		**2.805**	**1.657**	**0.329**	**0.204**	**0.005**
		PM=0		2.812 (0.1543)	1.642 (0.1412)	0.339 (0.0528)	0.202 (0.0347)	0.004 (0.0009)
		PM ∝ 1/Mean	With 0	1.412 (0.0948)	1.149 (0.0698)	0.975 (0.0548)	0.832 (0.0631)	0.631 (0.0747)
			Mean	1.700 (0.1178)	1.278 (0.1190)	0.905 (0.0812)	0.698 (0.0846)	0.419 (0.0743)
			EM	2.816 (0.1593)	1.645 (0.1515)	0.331 (0.0658)	0.186 (0.0433)	0.023 (0.011)
		PM∝ Mean	With 0	1.375 (0.0983)	1.152 (0.0636)	0.975 (0.0540)	0.829 (0.0615)	0.669 (0.0626)
			Mean	1.911 (0.1800)	1.277 (0.1190)	0.816 (0.0991)	0.612 (0.0819)	0.385 (0.0891)
			EM	2.811 (0.1661)	1.644 (0.1549)	0.342 (0.0683)	0.186 (0.0443)	0.018 (0.0102)

Table 5.12. First 5 eigenvalues for the simulated data (means of 100) in simulation II
n =100, p=10; S.E. in ()

p	Eigenvalues (ordered)			1	2	3	4	5
10	Low Noise	**True**		**5.487**	**3.618**	**0.304**	**0.207**	**0.135**
		PM=0		5.589 (0.1407)	3.508 (0.1349)	0.326 (0.0459)	0.208 (0.0340)	0.135 (0.0219)
		PM∝ 1/Mean	With 0	1.586 (0.1116)	1.384 (0.0634)	1.239 (0.0555)	1.120 (0.0491)	1.020 (0.0411)
			Mean	2.546 (0.2497)	2.059 (0.2042)	1.116 (0.1071)	0.933 (0.0783)	0.803 (0.0689)
			EM	5.591 (0.1544)	3.507 (0.1449)	0.334 (0.0476)	0.204 (0.0380)	0.132 (0.0237)
		PM∝Mean	With 0	1.686 (0.1460)	1.393 (0.0731)	1.240 (0.0569)	1.114 (0.0433)	1.002 (0.0545)
			Mean	3.696 (0.2799)	1.342 (0.1003)	1.095 (0.0680)	0.950 (0.0525)	0.807 (0.0612)
			EM	5.597 (0.1417)	3.502 (0.1422)	0.324 (0.0372)	0.202 (0.0372)	0.133 (0.0217)
	High Noise	**True**		**6.724**	**2.385**	**0.300**	**0.209**	**0.131**
		PM=0		6.737 (0.3694)	2.366 (0.3209)	0.322 (0.0473)	0.209 (0.0345)	0.132 (0.0221)
		PM ∝1/Mean	With 0	1.595 (0.1146)	1.388 (0.0658)	1.240 (0.0561)	1.117 (0.0503)	1.016 (0.0422)
			Mean	2.988 (0.3549)	1.568 (0.1720)	1.121 (0.0949)	0.943 (0.0756)	0.812 (0.0685)
			EM	6.729 (0.3814)	2.374 (0.3303)	0.329 (0.0497)	0.207 (0.0389)	0.132 (0.0249)
		PM∝Mean	With 0	1.693 (0.1471)	1.394 (0.0746)	1.240 (0.0566)	1.114 (0.0436)	0.999 (0.0525)
			Mean	3.762 (0.2875)	1.297 (0.0904)	1.083 (0.0627)	0.949 (0.0505)	0.807 (0.0637)
			EM	6.725 (0.3799)	2.378 (0.3265)	0.318 (0.0665)	0.203 (0.0378)	0.132 (0.0227)

Table 5.13. Eigenvalues for the simulated data (means of 100) in simulation II
n=40, p=5; S.E. in ()

p	Eigenvalues (ordered)			1	2	3	4	5
5	Low Noise	**True**		**2.613**	**1.791**	**0.318**	**0.201**	**0.077**
		PM=0		2.721 (0.1757)	1.700 (0.1701)	0.334 (0.0827)	0.175 (0.0442)	0.071 (0.0220)
		PM ∝1/Mean	With 0	1.485 (0.1331)	1.186 (0.0825)	0.970 (0.0763)	0.778 (0.0890)	0.582 (0.0820)
			Mean	1.652 (0.1699)	1.255 (0.1206)	0.945 (0.0876)	0.682 (0.1041)	0.465 (0.1114)
			EM	2.723 (0.2318)	1.687 (0.2002)	0.370 (0.1250)	0.162 (0.0542)	0.058 (0.0329)
		PM∝Mean	With 0	1.468 (0.1304)	1.182 (0.0823)	0.990 (0.0616)	0.779 (0.0862)	0.582 (0.0957)
			Mean	1.840 (0.247)	1.184 (0.1148)	0.920 (0.0946)	0.668 (0.1242)	0.387 (0.1365)
			EM	2.751 (0.1997)	1.674 (0.2042)	0.376 (0.1213)	0.147 (0.0612)	0.052 (0.0273)
	High Noise	**True**		**2.805**	**1.657**	**0.329**	**0.204**	**0.005**
		PM=0		2.825 (0.2284)	1.629 (0.2197)	0.356 (0.0863)	0.185 (0.0475)	0.005 (0.0014)
		PM ∝ 1/Mean	With 0	1.585 (0.1636)	1.188 (0.0898)	0.965 (0.0687)	0.745 (0.0915)	0.517 (0.0951)
			Mean	1.784 (0.1864)	1.277 (0.1345)	0.934 (0.1040)	0.636 (0.1189)	0.368 (0.0968)
			EM	2.807 (0.2672)	1.630 (0.2385)	0.378 (0.1409)	0.161 (0.0692)	0.024 (0.0158)
		PM∝ Mean	With 0	1.538 (0.1431)	1.192 (0.0909)	0.966 (0.0783)	0.766 (0.0857)	0.538 (0.0906)
			Mean	1.933 (0.2392)	1.280 (0.1591)	0.860 (0.1304)	0.588 (0.1182)	0.338 (0.1141)
			EM	2.844 (0.2334)	1.613 (0.2351)	0.375 (0.1192)	0.152 (0.0595)	0.017 (0.0127)

Table 5.14. First 5 eigenvalues for the simulated data (means of 100) in simulation II n =40, p=10; S.E. in ()

p	Eigenvalues (ordered)			1	2	3	4	5
10	Low Noise	**True**		**5.487**	**3.618**	**0.304**	**0.207**	**0.135**
		PM=0		5.699 (0.2856)	3.404 (0.2877)	0.353 (0.0755)	0.204 (0.0441)	0.131 (0.0315)
		PM∝ 1/Mean	With 0	1.926 (0.1484)	1.585 (0.1051)	1.369 (0.0789)	1.190 (0.0643)	1.008 (0.0634)
			Mean	2.755 (0.3478)	2.040 (0.2782)	1.292 (0.1532)	1.023 (0.1141)	0.836 (0.1029)
			EM	5.649 (0.3486)	3.413 (0.2984)	0.397 (0.1247)	0.225 (0.0585)	0.133 (0.0330)
		PM∝Mean	With 0	2.001 (0.1916)	1.616 (0.1148)	1.375 (0.0882)	1.161 (0.0856)	0.997 (0.0712)
			Mean	3.708 (0.4209)	1.522 (0.1649)	1.210 (0.1015)	0.977 (0.0868)	0.793 (0.0984)
			EM	5.725 (0.2975)	3.349 (0.3235)	0.391 (0.1363)	0.209 (0.0474)	0.135 (0.0326)
	High Noise	**True**		**6.724**	**2.385**	**0.300**	**0.209**	**0.131**
		PM=0		6.744 (0.5455)	2.367 (0.4810)	0.347 (0.0764)	0.203 (0.0444)	0.130 (0.0323)
		PM ∝1/Mean	With 0	1.932 (0.1470)	1.586 (0.1113)	1.370 (0.0807)	1.190 (0.0667)	1.008 (0.0663)
			Mean	3.079 (0.4701)	1.742 (0.2546)	1.272 (0.1332)	1.025 (0.1061)	0.834 (0.1005)
			EM	6.630 (0.6349)	2.400 (0.4783)	0.398 (0.1386)	0.239 (0.0903)	0.146 (0.0501)
		PM∝Mean	With 0	2.009 (0.1852)	1.620 (0.1144)	1.378 (0.0881)	1.160 (0.0864)	0.995 (0.0715)
			Mean	3.768 (0.4324)	1.511 (0.1606)	1.209 (0.1050)	0.967 (0.0892)	0.779 (0.0940)
			EM	6.705 (0.5507)	2.356 (0.4967)	0.406 (0.1537)	0.208 (0.0516)	0.136 (0.0336)

Table 5.15. The percentage of variation explained by the first two eigenvalues in simulation study II, n =100

	EM (re-constructed X)		Mean (re-constructed X)		With 0 (Y)		No mortality (complete X)	
Run Number	first eigen-value	second eigen-value	first eigen-value	second eigen-value	first eigen-value	second eigen-value	first eigen-value	second eigen-value
1	67.3%	23.8%	37.6%	13.0%	16.9%	13.9%	67.4%	23.7%
2	67.3%	23.7%	29.9%	15.7%	16.0%	13.9%	67.4%	23.7%
3	56.0%	35.0%	37.0%	13.4%	16.9%	13.9%	55.9%	35.1%
4	55.9%	35.1%	25.5%	20.6%	15.9%	13.8%	55.9%	35.1%
5	56.2%	32.9%	38.2%	25.5%	27.5%	23.0%	56.2%	32.8%
6	56.3%	32.9%	34%	25.6%	28.2%	23.0%	56.2%	32.8%
7	53.2%	35.1%	36.3%	22.6%	26.3%	22.4%	53.0%	35.3%
8	53.0%	35.2%	30.4%	25.0%	26.2%	22.5%	53.0%	35.3%

Table 5.16. The percentage of variation explained by the first two eigenvalues in simulation study II, n =40

	EM (re-constructed X)		Mean (re-constructed X)		With 0 (Y)		No mortality (complete X)	
Run Number	first eigen-value	second eigen-value	first eigen-value	second eigen-value	first eigen-value	second eigen-value	first eigen-value	second eigen-value
1	67.1%	23.6%	37.7%	15.1%	20.1%	16.2%	67.4%	23.7%
2	66.3%	24.0%	30.8%	17.4%	19.3%	15.9%	67.4%	23.7%
3	57.3%	33.5%	37.1%	15.2%	20.0%	16.2%	57.0%	34.0%
4	56.5%	34.1%	27.6%	20.4%	19.3%	15.9%	57.0%	34.0%
5	56.9%	32.3%	38.7%	25.6%	30.8%	23.8%	56.5%	32.6%
6	56.1%	32.6%	35.7%	25.5%	31.7%	23.8%	56.5%	32.6%
7	55.0%	33.5%	36.8%	23.7%	29.7%	23.6%	54.4%	34.0%
8	54.5%	33.7%	33.0%	25.1%	29.7%	23.7%	54.4%	34.0%

5.5 Conclusion

The problem of zero-heavy data often exists in environmental studies in which the growth or reproduction rates of multi-species are measured at numerous sites. This gives rise to multivariate data, and different multivariate methods can be applied to analyze the data. Since the inter-relationship between different species is of major interest to researchers, the focus of the study is always on this information. Principal component analysis is the most popular technique used to assess this information through the analysis of the correlation structure of the data. In the case where mortality influences the variables of interests, and when mortality is not the subject of interests, the use of the mixture approach can be applied to recover the information of the correlation structure. In order to investigate the effect of zeroes on multi-variate data, simulation studies on principal component analyses are performed. Two sets of simulations are done based on the relationship of the mortality probability and the mean of the observed data. The two cases are:

- constant mortality probability which does not depend on the mean of the non-zero data, and
- mortality probability either directly proportional or inversely proportional to mean of the non-zero data.

Many different estimation methods can be used to recover the correlation structure of the non-zero data. Only the mean substitution method and the EM algorithm are used in the simulation. The EM algorithm outperforms the mean method consistently, and is able to recover the correlation structure. However, the mean method is much easier to use.