BIAS IN PRINCIPAL COMPONENTS ANALYSIS DUE TO CORRELATED OBSERVATIONS

Hong Jiang

Kent M. Eskridge

BIAS IN PRINCIPAL COMPONENTS ANALYSIS DUE TO CORRELATED OBSERVATIONS

Hong Jiang and Kent M. Eskridge
Department of Biometry, University of Nebraska-Lincoln

ABSTRACT

A common practice in many scientific disciplines is to take measurements on several different variables on each unit from a designed experiment. This practice is cost efficient and results in data that may be analyzed using multivariate statistical methods. Usually, principal components analysis (PCA) is conducted by decomposing the covariance matrix of the several dependent variables using eigenanalysis without accounting for possible correlations among the observations. To evaluate how correlated observations bias PCA results, we used algebraic derivation and simulation for several different types of correlation structures. Our results indicated that sampling error generally had a much larger impact on the bias of PCA results than correlation between the observations. If we ignore the sampling error and there are no time trends or treatment effects, the PC's and the percent variance explained by a PC is not affected by correlated observations, however the eigenvalues are biased. If the sampling error is considered, for moderate sized correlations between observations and reasonably sized designs, bias was generally small enough to ignore for the first PC, otherwise SAS PROC MIXED may be used to easily correct for correlated observations, resulting in less bias in the PCA results.

1. INTRODUCTION

In most experiments in science and engineering, multiple variables are measured on each experimental unit. Typically experimental designs are the RCBD (randomized complete block designs) and the CRD (completely randomized designs), where single or repeated measurements are taken on each experimental unit for each of different variables.

In the analysis of such experiments, principal components analysis (PCA) is a useful multivariate method for understanding the nature of association among the variables. One of the goals of PCA is to reduce the dimension of the data from the total number of observed variables to a few meaningful "new " variables called principal components (PC's), that reduce the complexity of problem and aid with describing and understanding variation in the data. PC's are "composite" variables that are explanatory combinations of the original variables' where the coefficients display how each of the original variables' affects the PC's response and where the relative size of the coefficients give meaning to the component. The first few PC's usually account for the most of the variation in the data in which case they can be used to summarize the data with little loss of information (Johson and Wichen,1998; Johson, 1998; Morrison, 1976).

An important assumption of the PCA method is that all observation vectors are independent. PCA method is conducted without accounting for possible correlation among the observations in most applications. However, observations from scientific experiments will

generally be correlated, either due to repeated measurement of the experimental units, or due to the nature of the experimental design . Repeated measurement of experimental units will cause a dependence (or serial correlation) between the measurements taken over time. Repeated measurements arise in many fields, and are more common than single measurements. For example, in longitudinal individual studies, experimental units are monitored successively over a period of time to record the changing pattern of the responses (Crowder, 1996) . Even when experimental units are only measured once, the nature of the design of the experiment will often cause observations to be correlated. For example, in a RCB, with blocks random, all observations on a variable within a block are equally correlated and the resulting correlation structure is compound symmetric (CS). The more effective the blocking the larger, the correlation among observations within a block (Lentner,1993).

Therefore, an important question arises: how does correlation among observations affect PCA results when the correlation is ignored? We algebraically developed equations to determine the bias of eigenvalues and eigenvectors due to first order serial correlation (AR(1)) and CS correlated observations when the true covariance matrix was known. To evaluate the bias of PCA results as affected by both correlated observations and sampling error, we simulated multivariate normal observations with AR(1) and CS correlation structures. Then we computed eigenvalues and PC's based on the correlated observations and compared the results with the true eigenvalues and PC's. We also demonstrated how to accommodate serial correlation and reduce the bias of the PCA results using PROC MIXED.

2. THEORY AND METHODS

In scientific experiments, data on two variables are usually set up as following,

$$\begin{matrix} \underline{Y}_1 & \underline{Z}_1 \\ \underline{Y}_2 & \underline{Z}_2 \\ \vdots & \vdots \\ \vdots & \vdots \\ \underline{Y}_S & \underline{Z}_S \end{matrix}$$

where $\underline{Y}_1, \ldots, \underline{Y}_s$ are n×1 observation vector of variable Y on s experimental units for a repeated measure design or s blocks for a block design and $\underline{Z}_1, \ldots, \underline{Z}_s$ are observation vectors of variable Z. Assume that subjects (or blocks) are independent and for each subject the between variables correlation matrix free serial correlation is:

$$\Sigma = \begin{bmatrix} \sigma_y^2 & \sigma_{yz} \\ \sigma_{yz} & \sigma_z^2 \end{bmatrix}$$

To understand how observations are correlated among observations within each variable and across the variables, it is important to re-express data matrix as a column vector: $(\underline{Y}_1` \underline{Y}_2` \cdots \underline{Y}_S` \underline{Z}_1` \underline{Z}_2` \cdots \underline{Z}_s`)`$ Then the covariance matrix of this vector allows one to see these different types of correlations. For example, in an experiment with 2 subjects, 3 repeated measures and 2 variables per subject, the re-expressed data matrix is

$(Y_{11}\ Y_{12}\ Y_{13}\ Y_{21}\ \cdots Z_{23})'$ or in vector notation $(\underline{Y}_1'\ \underline{Y}_2'\ \underline{Z}_1'\ \underline{Z}_2')'$ with covariance matrix

$$\mathbf{V} = \begin{pmatrix} V_y & 0 & V_{yz} & 0 \\ 0 & V_y & 0 & V_{yz} \\ V_{yz} & 0 & V_z & 0 \\ 0 & V_{yz} & 0 & V_z \end{pmatrix} \qquad (1)$$

where $\mathbf{V}_y$ is a 3×3 covariance matrix within subject across observation; $\mathbf{V}_{yz}$ is a 3×3 covariance matrix within subject across variables; and $\mathbf{V}_z$ is defined similarly to $\mathbf{V}_y$. In this study, we assume $\mathbf{V}_y=\sigma_y^2\times\mathbf{R}$, $\mathbf{V}_z=\sigma_z^2\times\mathbf{R}$, $\mathbf{V}_{yz}=\sigma_{yz}\times\mathbf{R}$ where $\mathbf{R}$ is the across observation, within subject correlation matrix. $\mathbf{R}$ may take several different forms depending on the structure of the data. $\mathbf{V}$ may be extended to any number of subjects (or blocks), variables and repeated measures.

Types of correlation structures across observation

In this study, we analyzed two types of correlation structures across observation: first order autoregressive serial correlation and compound symmetric.

First order autoregressive serial correlation (AR(1))

Repeated measurements arise in many diverse fields. The term of repeated measure refers to situations where the same characteristic is observed, on the same experimental unit at different times. This means that when observations are made over time, the effect of the disturbance occurring at one period carries over into another period. For the AR(1) structure, the model is $y=x\beta+\varepsilon_t$; $\varepsilon_t=\rho\varepsilon_{t-1}+u_t$ where u_t is a normally and independently distributed random variable with mean zero and variance σ_u^2 and it is assumed to be independent of ε_{t-i}. It can be shown that $E(\varepsilon_t)=0$, $Var(\varepsilon_t)=\sigma_u^2/(1-\rho^2)=\sigma^2$, and $Cov(\varepsilon_t, \varepsilon_{t-i})=\rho^i\sigma^2$ for $i<t$, $Cov(\varepsilon_t, \varepsilon_{t-i})=\rho^i\sigma^2$ indicates that the greater the number of periods between two disturbances is, the smaller their covariance is (Kmenta,1971; Chatfield,1999).For example, with 2 subjects, 3 repeated measures and 2 variables, the $\mathbf{V}_y$ in covariance matrix $\mathbf{V}$ in (1) is $\mathbf{V}_y=\sigma_y^2\times\mathbf{R}$, where

$$\mathbf{R} = \begin{pmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{pmatrix}$$

In this study, we assumed the across variables covariance matrix, $\mathbf{V}_{yz}$ in (1), can be expressed in a similar manner, $\mathbf{V}_{yz}=\sigma_{yz}\times\mathbf{R}$

Compound symmetric correlation structure (CS)

A correlation matrix among observations is said to possess compound symmetry (CS) when it can be written in the form

$$\mathbf{R} = \begin{pmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho & \rho & \cdots & \rho & 1 \end{pmatrix}$$

This is a reasonable covariance structure for a single variable for many designed experiments (Morrison,1976). For example, consider the RCB design with random blocks.
The model is $Y_{ij}=\mu+\beta_j+\tau_i+\varepsilon_{ij}$; where $E(y_{ij})=\mu+\tau_i$, $V(y)=\sigma_\beta^2+\sigma_\varepsilon$ and $Cov=(y_{ij,}\, y_{i`j})=\sigma_\beta^2$ $i\neq i`$, and we define $\rho=\sigma_\beta^2/(\sigma_\beta^2+\sigma_\varepsilon^2)$, so $Corr(y_{ij,}\, y_{i`j})=\rho$. In this case the above covariance matrix **R** holds for a single block. In the example with 2 blocks, 3 treatments and 2 variables $\mathbf{V}_y=\sigma_y^2\times\mathbf{R}$, and $\mathbf{V}_{yz}=\sigma_{yz}\times\mathbf{R}$. Many other types of designed experiments result in similar but more complex covariance matrices.

Principal Component Analysis (PCA)

PCA is one of the most widely used methods in multivariate analysis. Principal components depend solely on the covariance matrix or correlation matrix of the original variables $x_1, x_2, \cdots, x_p$. The i^{th} PC is defined as the linear combination PC_i,

$$PC_i = e_{i1}x_1 + e_{i2}x_1 + \cdots + e_{ip}x_p \qquad i=1, 2, \cdots, p$$

where $\underline{e}_i$ is called an "eigenvector" having coefficients e_{ij}. The coefficient subscript i refers to the eigenvector index and j refers to the original variable (x_i) index. The eigenvectors and eigenvalues are obtained by performing eigenanalysis on the correlation (or covariance) matrix. The variance of each principle component PC_i is the eigenvalue, denoted by L_i, i=1, 2, ⋯, p, of the i^{th} eigenvector. This can be shown by noting that $Var(PC_i)=\underline{e}_i`\Sigma\underline{e}_i=L_i$ . In addition the covariance between any two different PC's is zero: $Cov(PC_i\ PC_k)=\underline{e}_i`\Sigma\underline{e}_k=0$. The proportion of total population variance explained by the kth principal component is $L_k/\sum^{\rho}_{i=1}(L_i)$. Where $\sum^{\rho}_{i=1}(L_i)$ is the total variance or the sum of all the variances of the original variables (Johnson and Wichern, 1998; Morrison, 1976).

Standard PCA implicitly assumes no covariance between any two observations, either within a variable or across variables. Accordingly, the covariance matrix **V** for two variables y and z in equation (1) is implicitly assumed to have **R=I**, i.e. all observations are uncorrelated.

Bias of eigenvalues, eigenvectors and percent variance explained by PCA

We consider two different situations in how precisely the covariance matrix is estimated:
(i) The population covariance matrix **V** is known, and the sampling error is not a consideration. In this case, we derive algebraic results based on expected SS and cross-products to see the relative influence correlation among observations on the bias of PCA results.
(ii) For small samples, we use simulation to assess the combined effects of sampling error in estimating **V** and correlated observations on bias of PCA results.

i. Algebraic method

In this section we will algebraically derive bias of eigenvalues, eigenvectors and percent variance explained by PC's with assumptions: (1) $\mathbf{V}$ has structure as in equation (1), (2) $\mathbf{V}$ is known for both AR(1) and CS correlation structure, and (3) there is no time trend for a repeated measures design or there are no treatment effects for a RCB design.

In many applications of PCA of experimental data, researchers estimate Σ with the sample covariance matrix, disregarding any correlation among observations. When there is correlation among observations, the sample covariance matrix is biased.
Assuming subjects are random, sample variance and covariance and their expected value E(.) are given as follows (Searle,1997):

$$S_y^2 = \frac{\Sigma\Sigma(y_{ij}-\bar{y}_{..})^2}{ns-1} = \frac{\mathbf{y'Ay}}{ns-1}; \qquad E(S_y^2) = \frac{E(\mathbf{y'Ay})}{ns-1} = \frac{tr(\mathbf{AV}_y)+\underline{\mu}_y'\mathbf{A}\underline{\mu}_y}{ns-1}$$

$$S_z^2 = \frac{\Sigma\Sigma(z_{ij}-\bar{z}_{..})^2}{ns-1} = \frac{\mathbf{z'Az}}{ns-1}; \qquad E(S_z^2) = \frac{E(\mathbf{z'Az})}{ns-1} = \frac{tr(\mathbf{AV}_z)+\underline{\mu}_z'\mathbf{A}\underline{\mu}_z}{ns-1}$$

$$S_{yz} = \frac{\Sigma\Sigma(y_{ij}-\bar{y}_{i.})(z_{ij}-\bar{z}_{i.})}{ns-1} = \frac{\mathbf{y`Ay}}{ns-1} \qquad E(S_{yz}) = \frac{E(\mathbf{y`Az})}{ns-1} = \frac{tr(\mathbf{AV}_{yz})+\underline{\mu}_y`\mathbf{A}\underline{\mu}_z}{ns-1}$$

where n is the number of repeated measures, s is the number of subjects (or blocks), and $\mathbf{A} = \mathbf{I}-\bar{\mathbf{J}} = \mathbf{I}-(1/ns)\mathbf{11}`$ which corrects for the mean.
Now,

$$tr(\mathbf{AV}_y) = tr(\mathbf{I}-\bar{\mathbf{J}})\mathbf{V}_y) = tr(\mathbf{I}-(1/ns)\mathbf{11}`)\mathbf{V}_y) = tr(\mathbf{V}_y) - tr(\frac{\mathbf{11}`\mathbf{V}_y}{ns}) =$$

$$\sigma_y^2 tr(\mathbf{R}) - \frac{\sigma_y^2}{ns} tr(\mathbf{11}`\mathbf{R}) = \sigma_y^2(ns - \sum_{i=1}^{n}\sum_{j=1}^{n} r_{ij})/n) \Rightarrow$$

$$E(S_y^2) = \sigma_y^2(1 - \frac{(\sum_{i=1}^{n}\sum_{j=1}^{n} r_{ij}) - n}{n(ns-1)}) + \underline{\mu}_y`\mathbf{A}\underline{\mu}_y/(ns-1) = c\sigma_y^2 + \underline{\mu}_y`\mathbf{A}\underline{\mu}_y/(ns-1)$$

where r_{ij} is the element of the R matrix.
In the same way, we can get

$$E(S_z^2) = c\sigma_z^2 + \underline{\mu}_z`\mathbf{A}\underline{\mu}_z/(ns-1) \qquad E(S_{yz}) = c\sigma_{yz} + \underline{\mu}_y`\mathbf{A}\underline{\mu}_z/(ns-1)$$

Similar result hold for any other pairs of variables.

If time trend or treatment effects are not removed, PCA results will be biased. If we want to decompose the error covariance matrix with PCA, we shoud first correct for the time trends or treatment effects.

In this study, we assumed that there were no time trends or treatment effects, so $\mu_{y1}=\mu_{y2}=\cdots\cdots=\mu_{yn}\Rightarrow \underline{\mu}_y' \mathbf{A}\, \underline{\mu}_y = 0 \Rightarrow E(\underline{y}'\mathbf{A}\underline{y}) = tr(\mathbf{A}\mathbf{V}_y)$, thus

$$E(\mathbf{S}) = c\begin{pmatrix} \sigma_y^2 & \sigma_{yz} \\ \sigma_{yz} & \sigma_z^2 \end{pmatrix}$$

The expectation of the sample covariance is a constant c times the true covariance matrix.When the covariance structure is AR(1) model, for p variables, n repeated measures, and s subjects, $c = (1 - \sum_{i=1}^{n-1} \frac{2(n-i)}{(sn-1)n}\rho^i)$ where c is a function of ρ, the number of subjects s and the number of time measurements n. When the data have CS covariance structure, $c = (1 - \frac{n(n-1)}{(sn-1)n}\rho)$.

When observations are not independent, the sample covariance matrix estimates $c\Sigma$ and so is biased. However, as the sample size s and/or n, the number of time measurements gets large, c approaches 1 and the sample covariance matrix is a consistent estimate of Σ.

We assume that L_i is an eigenvalue of the true covariance matrix Σ and $\underline{e}_i$ is the eigenvector corresponding to L_i. Yet LC_i be the eigenvalue of the covariance matrix $c\Sigma$. Then $LC_i=cL_i$.Thus the eigenvalue is biased by the factor c. The bias as a proportion of the true eigenvalue is $LC_i-L_i)/L_i=(cL_i-L_i)/L_i=c-1$. Thus the estimated eigenvalues are biased by the proportion c-1. Now note that $L_1/\Sigma(L_i)$=true proportion of total variance explained of the 1st PC and $(LC_1/\Sigma(LC_i)$= estimated proportion of total variance explained of 1st PC. Then $LC_1/\Sigma(LC_i)=cL_1/\Sigma(cL_i)= L_1/\Sigma(L_i)$ and so the proportion of total variance explained by any PC is unbiased. Now let $\underline{e}_{ci}$ be the eigenvectors corresponding to LC_i and using the definition of eigenvectors $c\Sigma\underline{e}_{ci}=cL_i\underline{e}_{ci} \Rightarrow \Sigma\underline{e}_{ci}=(L_i)\underline{e}_{ci.}$ But this last equality holds for the true eigenvector So $\underline{e}_i=\underline{e}_{ci}$ and all eigenvectors are thus unbiased for all the PC's. These results hold for any balance linear mixed model as long as the sampling error is zero and there are no time trends or treatment effects.

ii. Simulation method

We used simulation to evaluate how bias of the PCA results were effected by sampling error in estimating the covariance matrix and by correlation among the observations. We used the SAS/MVN macro and SAS/IML to generate multivariate normal data (SAS MVN Macro) and SAS/IML to obtain the eigenvalues and eigenvectors of the sample covariance matrix and to evaluate the bias in the principle components results due to the repeated measures (SAS Institute Inc, 1990). We simulated observations for CS and AR(1) among observation covariance structure for values of ρ with 0.2, 0.5,or 0.9, for correlation among variables ρ_{yz} as 0.3, 0.7, and the number of repeated measures n equal to 3 and 10. 500 samples were generated for each case.

3. RESULTS AND DISCUSSION

Algebraic results

Based on the assumptions of **V** as in equation (1), with no trend or treatment effects, and no sampling error, the PC's are not affected by correlated observations. In addition, the percent variance explained by a PC is not affected by correlated observations. However, the proportional bias of the eigenvlaues is c-1. This bias can be severe when ρ is large, or the number of subjects s (or blocks) and the number of repeated measures n (or treatment) are small.

Simulation results

Figures 1a and 1b displayed the simulated proportional bias of the first and second eigenvalues that include sampling error for the AR(1) model with ρ_{yz}=0.3, 2 variables and 3 repeated measures. When s is large, the bias asymptotically approaches zero. The simulation results suggested that the number of subjects (or blocks) should be greater than 30 to keep bias of the eigenvalues on the two PC's below 15%. Comparing Figures 1a and 1b where the correlation among the variables is 0.3 with Figures 2a and 2b where the correlation among the variables is 0.7, we find the curves are somewhat different for PC1 and similar for PC2. Generally, it appears that the larger the correlation among the variables, the smaller the bias of the eigenvalues.

Figures 3a and 3b display the simulated proportional bias of the variance explained by the PC's that include sampling error for the AR(1) model with 2 variables, 3 repeated measures, and ρ_{yz}=0.3. When the number of subjects is large (>30), the bias is less than 15% and asymptotically approaches zero as the number of subjects become very large. But when s is small, bias can be large as compared to the algebraic no sampling error case where the bias is zero. The percent variance explained by the PC's is also affected by the size of correlation between variables where the larger the correlation among variables, the smaller the bias (Jiang, 2000).

Figures 4a and 4b show the proportional bias of the loadings for PC1. We see that the bias trends are similar for both loadings. So interpretation of loadings appears to be similar to interpretation of the true loadings.

Figures 5 –7 display the bias of PCA for 4 variables and 3 repeated measures when sampling error is considered. From Figures 5a and 5b, we can see that the first two eigenvalues all have positive bias. When the number of subjects is greater than 30 or the serial correlation is not very large, bias is less than 10% for the first eigenvalue. But there is a large positive bias (over 25%) for the second eigenvalue with s=30. We need to be careful about the second eigenvalue even though the number of subject is at least 30.

From Figure 6a, bias of the percent variance explained by PC1 is generally less than 15%, when number of subjects is greater than 10 . However, from Figure 6b, bias of the percent variance explained by PC2 can be large (>25%), even with 30 subjects.

Figures 7a, 7b, 7c, 7d show the proportional bias of the loadings of PC1. All four loadings have negative bias and similar bias trends, so interpretation of PC1 is relatively

unaffected by sampling error. On the loading for the other PC's, the bias trends are not similar, and consequently, interpretationes can be quite biased. Bias for the proportion of variance explained by the first PC is rather small but can be large for the other PC's Very similar results are obtained using the CS correlation structure.(Jiang, 2000).

4. DATA ANALYSIS

In small samples (s<30), it is likely useful to use an estimation procedure to obtain variance component estimates after accounting for serial correlation. The covariance matrix using these variance and covariance estimates may then be decomposed using PCA In this section we demonstrate how to use PROC MIXED to get these variance estimates, and compare the PCA results with those based on the sample covariance matrix.

We applied this approach to data from 22 children who were measured monthly over eighteen monthes for blood lead, urine lead, food lead, and lead on the childrens' hand denoted PB_BLOOD, UA, PB_FOOD and HDWP (Stanek, et al, 1998). We used the model is $y=\mu+s_i+e_{ij}$, where $s_i=i^{th}$ subject effect and $e_{ij} \sim mvn(0, \Sigma)$. PROC MIXED (Littell, et al, 1996) was used to estimate the covariance matrix after removing the effects of serial correlation. For these repeated measures data, we used the UN@AR(1) covariance matrix in MIXED. This covariance matrix is the same as the **V** matrix we defined as above, that is, all variables have same across observation correlation structure, both within and among variables, where **R** is from the AR(1) serial correlation structure.

To use PROC MIXED, all variables were stacked into a new single variable Y with the following SAS statements:

```
Proc Mixed;
Class sub var visit ;  /* sub is subject number, var identifies the variable,
                          visit is repeated measurement #  */
Model Y=var/noint;   /* Y is the stacked new variable */
Repeated var visit /subject=sub type=UN@AR(1);
```

Using this program we obtained the covariance matrix after removing the effect of serial correlation:

Covariance Parameter Estimates (REML)

Cov	Parm	Subject	Estimate	Variables
VAR	UN(1,1)	SUBJNO	20.37212817	HDWP
	UN(2,1)	SUBJNO	1.16684061	
	UN(2,2)	SUBJNO	12.01942312	PB_FOOD
	UN(3,1)	SUBJNO	1.38920152	
	UN(3,2)	SUBJNO	0.23577516	
	UN(3,3)	SUBJNO	3.38313055	UA
	UN(4,1)	SUBJNO	2.26274525	
	UN(4,2)	SUBJNO	0.39611874	
	UN(4,3)	SUBJNO	2.8453265	
	UN(4,4)	SUBJNO	8.8123017	PB_BLOOD
Serial	VISIT	AR(1)	SUBJ	0.40341359
Correlation				

Then we used this covariance matrix, which is not influenced by serial correlation, to obtain eigenvalues and eigenvectors of the correlation matrix, using Proc Princomp. We then compared these PCA results with those based on the sample correlation which contained the effect of serial correlation.

Eigenvalues of the Correlation Matrix after removing effects of serial correlation

	Eigenvalue	Difference	Proportion	Cumulative
PRIN1	1.62277	0.608803	0.405693	0.40569
PRIN2	1.01397	0.129596	0.253492	0.65919
PRIN3	0.88437	0.405485	0.221093	0.88028
PRIN4	0.47889	.	0.119722	1

Eigenvalues using sample correlation matrix.

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
PRIN1	1.66388	0.584978	0.41597	0.41597
PRIN2	1.0789	0.330903	0.269726	0.6857
PRIN3	0.748	0.238785	0.187	0.8727
PRIN4	0.50922	.	0.127304	1

Eigenvectors of the Correlation Matrix after removing effects of serial correlation

	Eigenvectors			
	PRIN1	PRIN2	PRIN3	PRIN4
HDWP	0.366749	0.354769	0.860017	0.002094
PB_FOOD	0.122927	0.897845	-0.422799	0.001815
UA	0.651739	-0.186182	-0.202848	0.706701
PB_BLD	0.652396	-0.182617	-0.201156	-0.7075

Eigenvectors on samplecorrelation matrix

	Eigenvectors			
	PRIN1	PRIN2	PRIN3	PRIN4
HDWP	0.41539	0.534728	-0.734024	0.052206
PB_FOOD	0.352991	0.64679	0.674355	0.048022
UA	0.60836	-0.332054	0.051238	-0.719036
PB_BLD	0.576842	-0.430661	0.061877	0.691343

Proportion bias of eigenvectors (assuming mixed results are unbiased)

	PRIN1	PRIN2	PRIN3	PRIN4
HDWP	0.1326274	0.5072569	-1.853499	23.934743
PB_FOOD	1.8715578	-0.27962	-2.594975	25.461639
UA	-0.066559	0.7834935	-1.252595	-2.017455
PB_BLD	-0.115811	1.3582782	-1.307609	-1.977154

Using Proc Mixed to correct for serial correlation in these data did not change the eigenvalues results much compared to the PCA results based on the sample covariance matrix. There appeared to be considerable bias in some of the loadings. PC1 uncorrected for serial correlation shows a much higher loading on food lead than PC1 on the results corrected for serial correlation.

5. CONCLUSIONS

This study evaluated the effects of correlation between observations, and the variables, and sampling error on the bias in PCA results. In general, sampling error had a much larger impact on the bias of PCA results than correlation among the observations.

In most cases evaluated, the bias of eigenvalues, The proportion of variance explained by the PC's and interpretation of the loadings was not severe for PC1 when there were at least 30 subjects. Bias for PC2 and higher could still be large even with more than 30 subjects.

As a general rule of thumb, if one is mostly interested in the first PC, then analysis of the sample covariance matrix, without accounting for the effects of serial correlation is probably acceptable if the experiment has more than 30 subjects for a repeated measures design, or 30 blocks for a simple block design, however, fewer subjects may be required with more repeated measures (n>30) (Jiang, 2000). If the number of subjects or blocks is smaller than 30, and/or the researcher is interested in PC's beyond the first, it may be better to first correct for the serial correlation, before PCA is conducted. We are currently investigating the reduction in bias using this approach.

REFERENCES

Chatfield, Chris, *The Analysis of Time Series, An Introduction*, 5th edition, Chapman & Hall, (1999).

Crowder, M. J. and Hand, D. J., *Analysis of Repeated Measures,* Chapman & Hall, (1996).

Jiang, Hong, Bias in Principal Components Analysis due to Correlated Observations, *M.S. Thesis*, The University of Nebraska-Lincoln, (2000).

Johnson, Richard A. and Wichern, Dean W., *Applied Multivariate Statistical Analsis*, Prentice Hall, (1998).

Johnson, Dallas E., *Applied Multivariate Methods for Data Analysts*, Brooks/Cole Publishing, (1998).

Kmenta, Jan, *Elements of Econometrics*, The Macmillan Company, (1971).

Lentner, Marvin and Bishop, Thomas, *Experimental Design and Analysis*, Valley Book, (1993).

Littell, Ramon C., Milliken, George A., Stroup, Walter W. and Wolfinger, Russell D., *SAS System for Mixed Models*, SAS Institute Inc., (1996).

Morrison, Donald F., *Multivariate Statistical Methods*, McGraw-Hill, (1976).

Otter, Pieter W. and Schuur, Jan F., "Principal Component Analysis in Multivariate Forecasting of Economic Time Series," *Time Series Analysis: Theory and Practice 1 (Proceedings of the International Conference held at Valencia, Spain),* (1982).

SAS Institute Inc., *SAS/IML Software: Usage and Reference,* Version 6, SAS Institute Inc., (1990).

SAS MVN MACRO: http://ftp.sas.com

Stanek, K. L., Manton, W. L., Angle, C. R., Eskridge, K. M., et al., "Lead consumption, anthropometric measurements, and nutrient intake of young children as determined from duplicate diet collections," *Journal of the American Dietetic Association* , 98(2): 155-159, (1998).

Searle, S. R., *Linear Models*, John Wiley & Sons, (1971).

Figure 1a. Proportional bias of first eigenvalue
(AR1 model, includes sampling error)

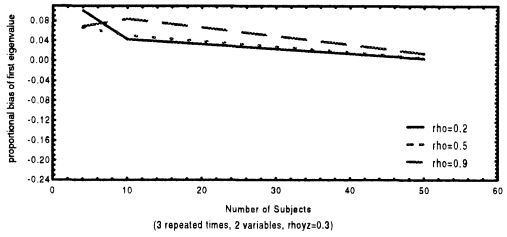

Figure 1b. Proportional bias of second eigenvalue
(AR(1) model includes sampling error)

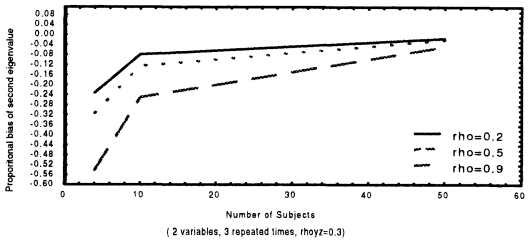

Figure 2a. Proportional bias of first eigenvalue
(AR1 model, includes sampling error)

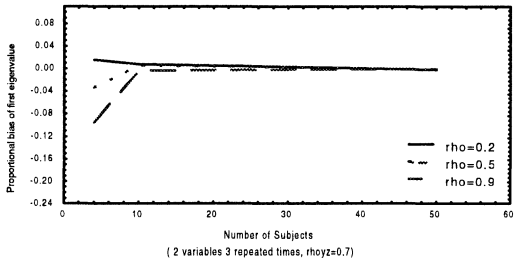

Figure 2b. Proportional bias of second eigenvalue
(AR1 model, includes sampling error)

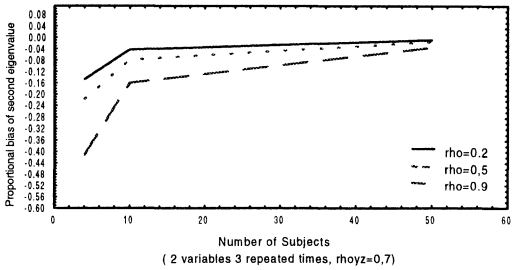

Figure 3a. Bias of percent variance explained of first PC
(AR(1) model includes sampling error)

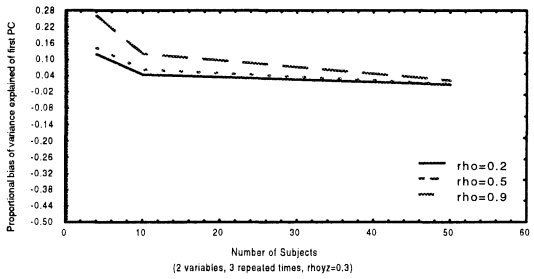

Figure 3b. Bias of percent variance explained of second PC
(AR(1) model, includes sampling error)

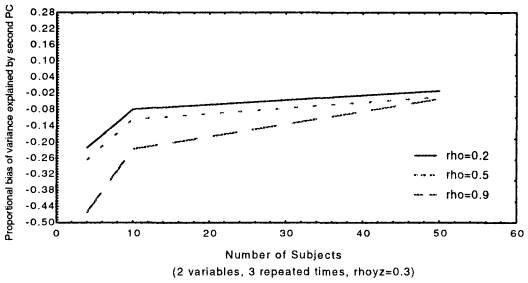

Figure 4a. Proportional bias of first loading of PC1
(AR(1) model, includes sampling error)

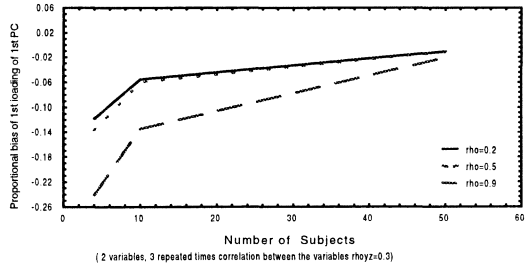

Figure 4b. Proportional bias of second loading of PC1
(AR(1) model, includes sampling error)

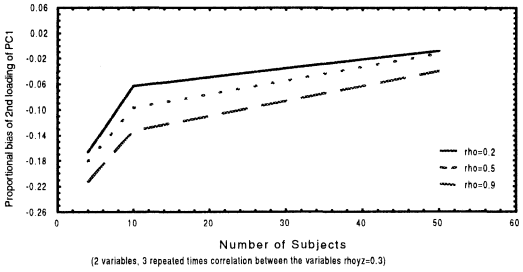

Figure 5a. Proportional bias of first eigenvalue
(AR(1) model, include simpling error)

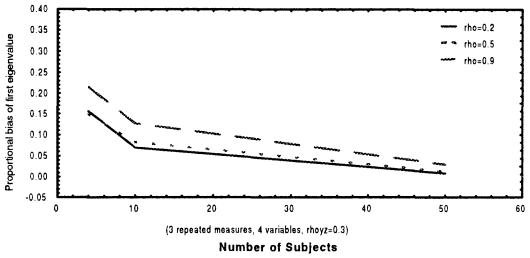

Figure 5b. Proportional bias of second eigenvalue
(AR(1) model, includ simping error)

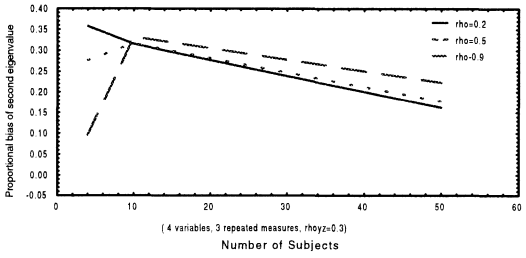

Figure 6a. Bias of percent variance explained of Ist PC
(AR(1) model include sampling error)

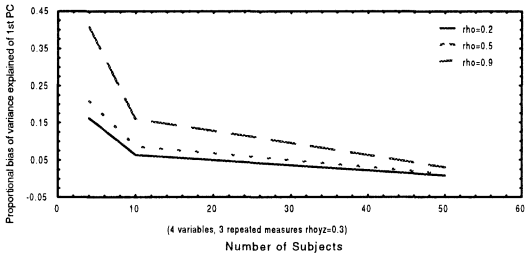

Figure 6b. Bias of percent variance explained of 2nd PC
(AR(1) model include sampling error)

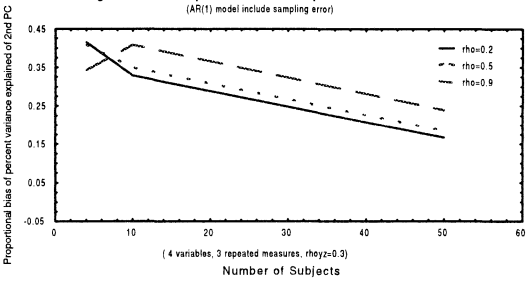

Figure 7a: Proportional bias of first loading of PC1
(AR(1) model, include sampling error)

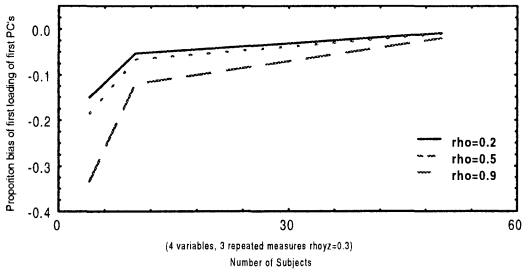

Figure 7b: Proportional bias of second loading of PC1
(AR(1) model, include sampling error)

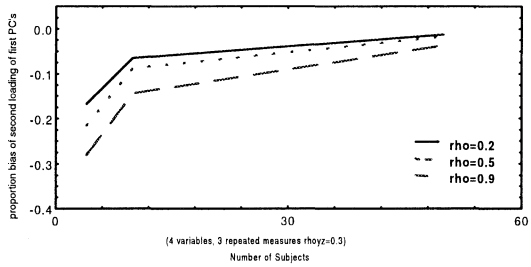

Figure 7c: Proportional bias of third loading of PC1
(AR(1) model, include sampling error)

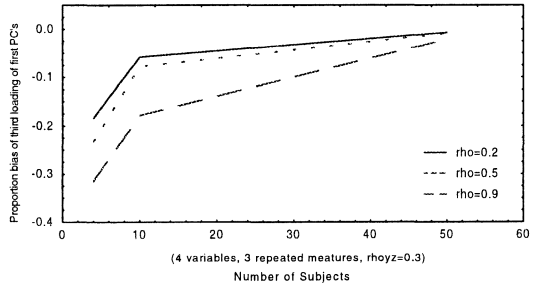

Figure 7d: Proportional bias of forth loading of PC1
(AR(1) model, include sampling error)