Controlling the Effect of Sample Design on Principal Components Analysis – A Simulation Study

Pedro J. Saavedra[1], Francine Barrington[2], Andrey Vinokurov[3]
[123]Macro International, 11875 Beltsville Dr, Calverton, MD, 20705

Abstract

Principal Components Analyses are typically conducted without taking into account the sampling design. Controlling for variables that are part of the sample design may affect the interrelationship of variables in a manner that is difficult to interpret. Controlling for clusters, strata or probabilities of selection may affect the substance or the stability of the results. In order to examine this issue, a data base of zip code areas was stratified and clustered by counties or combinations of adjacent counties. PCAs were conducted for several random samples, for cluster samples, and for several stratified designs. Clustering diminished the ability of the sample to reproduce the population PCA, with adjustments producing mixed results. Stratification with a similar number of units selected per stratum and proper weighting led to results at times better than random sampling in spite of inequality of weights. Simulations also indicated that stability of results is in part a function of sample design.

Key Words: Stratified sample, cluster sample, eigenvalues, phi coefficient

1. Introduction

Factor Analysis and Principal Components Analysis (PCA) originated as a means of discerning simple structure from the interrelationship of variables. As such, while the stability of the results has always been of concern, confidence intervals and tests of significance have been of secondary importance to the practitioners. In addition, the two techniques have been primarily used by psychometricians who have used a sample of convenience or even an entire population, and thus any effects of sample design have not been given sufficient consideration.

When dealing with survey data, if one wishes to implement linear regression there are two approaches: 1) using a procedure designed for survey data (such as PROC SURVEYREG in SAS) to define the weights, clusters and strata and have the regression procedure take them into account, or 2) include the variables designed to create clusters and/or strata and the weights into the regression model. These strategies are necessary for conducting regression analyses on the survey data, because considerations of degrees of freedom and confidence intervals enter into the interpretation of the results. But in conducting a PCA the main objective is to detect structure in the relationships between variables. And if one has a complex sample design one would like to obtain the same PCA results (or as close an approximation as possible) as if one had been able to conduct the analysis with the entire population. So the issue is how to achieve this when one is using a clustered or a stratified sample.

Previous research conducted by Skinner, Holmes, and Smith (1986) performed an analysis regarding the effect of sample design on Principal Component Analysis. In their study they examined the effects of sample design on both principal component analysis as well as the use of alternative maximum likelihood and probability-weighted procedures, conducting a simulation study of the properties of alternative estimators. Their samples consisted of a random sample, as well as six stratified samples: one stratified sample was proportionally allocated, while the others were of varying increasing and u-shaped allocations. Skinner, Holmes, and Smith (1986) found that estimators showed biased for non-self-weighting sample designs.

This study attempts to examine other types of samples, including clustered samples, as well as samples that are both weighted and un-weighted. In addition, this study examines the stability of the principal components, as well as principal components beyond the first two components. Our study found that several adjustment methods were counter-productive. This does not mean that geographical clusters can be ignored. Simulations indicated that there may be greater instability of results when a cluster sample is used instead of a simple random sample. But the results may still represent a good enough approximation.

Section 2 will discuss the dataset and samples, while section 3 will discuss the population principal components analysis. Section 4 will address the stability of the PCAs. Section 5 discuss the simulation study and present the results. Lastly, section 6 will conclude.

2. Dataset and Samples

2.1 The Database

In order to examine the effects of clustering on the PCA results we used a data set obtained from the 1990 Census. The data points were zip code areas. Twenty four variables were used. Those variables which would be artifactually correlated with the population (such as population in any particular ethnic group or number of households) were divided by the population in the zip code area. Zip code areas with very few or no residents (such as zip codes assigned to a government building) were omitted. The data base had variables related to population, housing values, density, area, proportion of minorities, average age, car ownership and similar variables. The substantive interpretation of the principal components is beyond the scope of this paper, though variables are identified by the name used in the data base.

2.2 The Samples

2.2.1 Random Samples and Other Designs

Random samples of 1000 zip code areas were drawn and principal components analysis was conducted for the population of zip code areas and for each random sample. In addition, cluster samples and stratified samples were also drawn, and principal components analysis was also conducted. However, on top of a standard PCA analysis one additional adjustment was tried for the cluster sample, and one of the stratified samples was analyzed with and without weights.

2.2.1 Clustering the Units and Sampling Clusters

The initial clusters were counties. A clustering algorithm was used which required a cluster to have at least 15 zip code areas. If a county was not large enough to create a cluster it was merged with the nearest county or cluster. The process continued until every cluster had at least 15 zip code areas. The frame had 24,954 zip code areas and 825 clusters. The cluster samples were selected with PPS (number of zip code areas in the cluster) using randomized systematic sampling with probability minimum replacement. A total of 100 clusters were drawn for each analysis, with ten units selected from each cluster.

There were two analytic schemes that attempted to control for design variables and sample clustering, but were found to be counter-productive. . The first sample was created using dummy variables for clusters using partial correlations. The second sample was created using standard scores within clusters. Both of these attempts destroyed the intercorrelations of the study variables, resulting in very different Principal Component solutions from each other and from the Principal Components Analysis that was run from the population. That effort was dropped altogether, as the objective of the study was to try to obtain solutions that were similar to principal components analysis conducted from the entire population. These two results will not be discussed further though they could be of interest if one wished to examined simple structure controlling for geographical proximity. Similar results could be obtained without sampling if one controlled for between cluster covariances.

There were two additional analytic schemes that yielded reasonable results with each cluster samples. The first simply ignored the clustering and treated the sample the same way as if it were a random sample. The second created ten classes, each including one unit per cluster, and then standardized the variables within class. The PCA was then conducted on the standardized variables.

2.2.2 Stratification and Stratified Samples

In stratified samples, strata were created using HUD regions. Two stratified sample designs were used. One sampled an equal number of units per region, while the other was created using a proportional number of units per region. Two analytic strategies were used with the first set, one using weights and one without weights. The second sampling scheme was self-weighting so only one analytic scheme was used.

2.2.3 The Samples

Six combinations of sampling and analytic schemes were used to compare to the population PCA. From the frame both simple random samples, cluster samples, and stratified samples, each with 1,000 zip code areas, were drawn. For each sampling scheme 10,000 different samples were drawn. The following are the six sampling and analytic schemes used in the analysis:

1) Random Sample of 1,000 units

2) Stratified sample using HUD regions – 1,000 sampled proportional to the number of units per region.

3) Stratified sample using HUD regions – 1,000 sampled an equal number of units per region. Unweighted PCA.

4) Stratified sample using HUD regions – 1,000 sampled an equal number of units per region. Weighted PCA.

5) Clustered sample – 100 clusters and ten units per cluster.

6) Clustered sample – 100 clusters and ten units per cluster. Ten adjustment classes selected, each with one unit from each cluster. Variables standardized by adjustment classes.

3. The Population Principal Components Analysis

The Population Principal Components Analysis yielded six eigenvalues greater than one. The six factors collectively explained 74.4 percent of the variance. A Scree test performed also suggested six factors. A varimax rotation yielded interpretable factors. Even though the analysis was a principal components analysis, it is common to treat PCAs as if they were factor analyses, to identify the number of meaningful components and to submit that number to a varimax rotation. For this reason we refer to the components as factors in the table, and we use a common criterion to determine the number of meaningful factors.

Table 1: Population Eigenvalues
(Eigenvalues greater than one outlined)

	Eigenvalue	*Difference*	*Proportion*	*Cumulative*
1.000	6.202	1.600	0.258	0.258
2.000	4.601	1.893	0.192	0.450
3.000	2.708	0.943	0.113	0.563
4.000	1.765	0.345	0.074	0.637
5.000	1.420	0.271	0.059	0.696
6.000	1.150	0.290	0.048	0.744
7.000	0.860	0.089	0.036	0.779
8.000	0.770	0.066	0.032	0.812

9.000	0.704	0.094	0.029	0.841
10.000	0.610	0.063	0.025	0.866
11.000	0.547	0.044	0.023	0.889
12.000	0.503	0.066	0.021	0.910
13.000	0.437	0.072	0.018	0.928
14.000	0.365	0.072	0.015	0.944
15.000	0.293	0.044	0.012	0.956
16.000	0.249	0.045	0.010	0.966
17.000	0.204	0.061	0.009	0.975
18.000	0.143	0.014	0.006	0.981
19.000	0.129	0.002	0.005	0.986
20.000	0.126	0.005	0.005	0.991
21.000	0.121	0.044	0.005	0.996
22.000	0.077	0.061	0.003	0.999
23.000	0.016	0.016	0.001	1.000
24.000	0.000		0.000	1.000

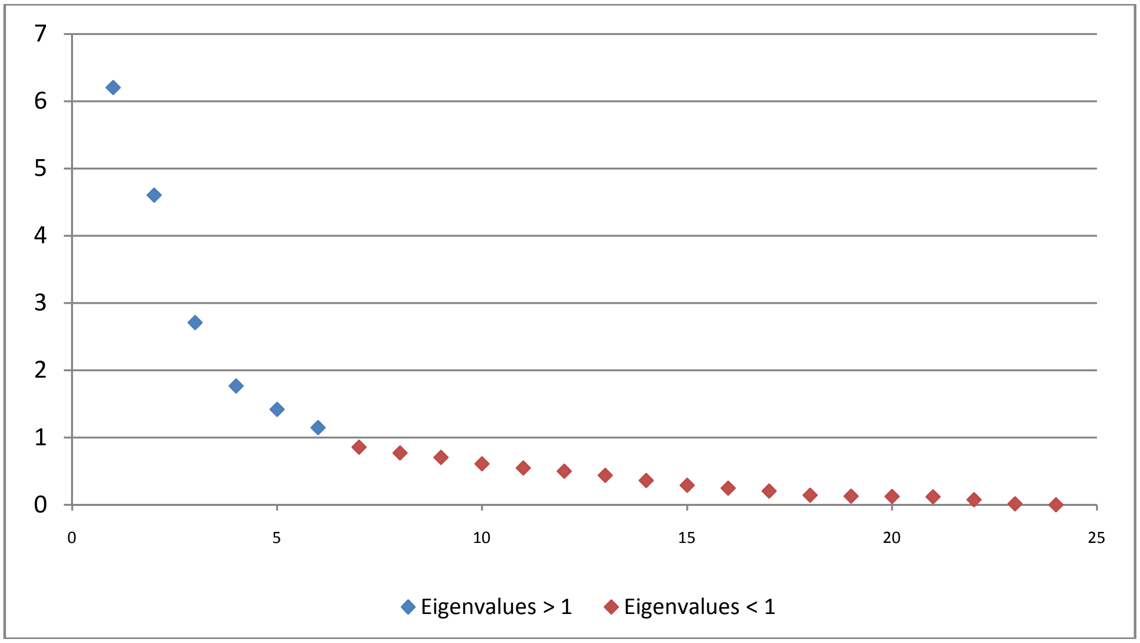

Figure 1: Graph of Population Eigenvalues

Table 2: Population Factor Pattern

	Factor1	*Factor2*	*Factor3*	*Factor4*	*Factor5*	*Factor6*
TOTPOP90	0.638	0.292	0.145	-0.136	-0.060	-0.084
WHITEPCT	-0.150	-0.822	0.156	0.116	0.264	-0.202
BLACKPCT	0.004	0.622	0.034	-0.359	-0.526	0.305
AMINDPCT	-0.155	0.346	-0.372	0.205	0.498	0.351
ASIANPCT	0.480	0.192	-0.022	0.335	-0.051	-0.181
HOUSPER	0.000	-0.092	0.894	0.002	0.320	0.208
SINGLE90	-0.733	0.104	0.241	0.109	-0.230	-0.103
OLDHSE90	-0.528	0.091	0.571	0.325	-0.115	-0.190
OCCUPER	0.000	-0.092	0.894	0.002	0.320	0.208

OWNERPCT	-0.486	-0.597	-0.212	0.276	-0.175	0.126
MEDVALUE	0.740	-0.085	0.080	0.447	-0.071	0.257
R499PCT	0.417	-0.128	-0.057	-0.558	0.253	-0.147
MEDRENT	0.863	-0.168	-0.013	0.257	-0.017	0.208
MEDINCOM	0.660	-0.458	-0.138	0.276	-0.152	0.307
URBANPOP	0.737	0.247	0.207	-0.161	0.000	-0.008
NATIBORN	-0.653	-0.382	-0.029	-0.377	-0.029	0.365
SAMEHOUS	-0.612	-0.113	0.067	0.378	-0.378	0.103
ENGLISH	-0.403	-0.516	0.183	-0.394	-0.190	0.383
LABOR	0.599	-0.414	-0.239	-0.200	0.004	0.215
UNEMPL	-0.287	0.692	-0.144	-0.045	0.152	-0.017
PUBTRANS	0.438	0.534	0.277	0.121	-0.225	0.118
NOPLUMB	-0.420	0.347	-0.305	0.281	0.341	0.312
NOPHONE	-0.499	0.643	-0.230	0.044	0.225	0.115
NOCAR	0.022	0.828	0.288	-0.029	-0.068	0.130

Variance Explained by Each Factor

Factor1	*Factor2*	*Factor3*	*Factor4*	*Factor5*	*Factor6*
6.202	4.601	2.708	1.765	1.420	1.150

Final Communality Estimates: Total = 17 0.846883

TOTPOP90	*WHITEPCT*	*BLACKPCT*	*AMINDPCT*	*ASIANPCT*	*HOUSPER*
0.542	0.847	0.886	0.696	0.415	0.954

Table 3: Population Rotated Factor Pattern

	Factor1	*Factor2*	*Factor3*	*Factor4*	*Factor5*	*Factor6*
TOTPOP90	0.197	0.370	-0.403	-0.395	0.082	-0.203
WHITEPCT	0.074	-0.828	0.201	0.060	0.219	-0.252
BLACKPCT	-0.096	0.911	0.158	0.020	-0.142	-0.042
AMINDPCT	0.011	0.007	-0.026	-0.049	-0.056	0.831
ASIANPCT	0.250	0.087	-0.582	-0.026	-0.057	-0.052
HOUSPER	0.006	-0.009	0.070	0.006	0.970	-0.090
SINGLE90	-0.517	0.041	0.214	0.597	0.097	-0.027
OLDHSE90	-0.415	-0.027	-0.072	0.624	0.433	-0.118
OCCUPER	0.006	-0.009	0.070	0.006	0.970	-0.090
OWNERPCT	0.123	-0.491	0.434	0.522	-0.210	-0.014
MEDVALUE	0.794	0.076	-0.418	-0.018	0.136	-0.049
R499PCT	0.012	-0.064	0.041	-0.739	-0.007	-0.197
MEDRENT	0.814	0.026	-0.371	-0.254	0.055	-0.118
MEDINCOM	0.888	-0.143	-0.088	-0.073	-0.084	-0.169
URBANPOP	0.295	0.367	-0.388	-0.475	0.178	-0.206
NATIBORN	-0.182	-0.147	0.878	0.140	0.019	0.060
SAMEHOUS	-0.114	-0.052	0.239	0.783	-0.050	0.000
ENGLISH	0.030	-0.086	0.839	0.110	0.153	-0.228
LABOR	0.613	-0.119	0.114	-0.463	-0.162	-0.175
UNEMPL	-0.466	0.373	-0.141	0.009	-0.091	0.473
PUBTRANS	0.190	0.619	-0.420	-0.004	0.187	-0.047
NOPLUMB	-0.139	0.017	0.056	0.224	-0.066	0.778
NOPHONE	-0.478	0.280	0.002	0.147	-0.107	0.665
NOCAR	-0.246	0.739	-0.271	0.035	0.242	0.226

4. Results of the Simulations

Six evaluation criteria were examined for each set of PCAs.

1) The number of eigenvalues greater than one.

2) The size of the first eigenvalue.

3) The sum of the first six eigenvalues.

4) The sum of the eigenvalues greater than one.

5) The absolute value of the phi coefficient of the first unrotated component and the first unrotated population component. (Phi= $\Sigma\ a_i b_i / \Sigma\ a_i^2\ \Sigma\ b_i^2$)

6) The absolute value of the phi coefficient of each subsequent unrotated component and the corresponding population component.

For the last criterion, in order to control for the possibility of two components switching order, the component that best reproduced the population component was assigned to it. This same concern led to avoidance of any examination of rotated components. Simple structure following rotation can be unstable, and two solutions may be similar, but the orthogonal transformation could be quite different.

Absolute values were taken because a principal component is defined up to a factor of -1. Thus, in some cases the coefficients of a component were close the negative of the coefficients for the same component in the population PCA. Simple t-tests were conducted to establish differences in means and variances across 10,000 samples for all six measures.

The objectives of the simulation study were to consider the effects of Clustering, Stratification and Weighting and reproduce with a sample the Population PCA. If design variables are related to the study variables, the objective is not to discover or control for the design variables, but to reproduce the correlation pattern found in the population. In order to explore this issue, six sets of 10,000 samples were treated as independent sets and examined.

4.1 The Results – Phi Coefficients

The stratified sample with the same number of units per stratum performed best, while the two clustered sampling methods performed worst. Adjustments to the clustering fared better (by a small but significant amount) for the first two components, but were counterproductive for the last four components. Random and stratified proportional were almost as good as the stratified weighted method. The stratified sampling method with the same number of units per stratum did better than the two clustered sampling methods, but worse than the random, the stratified proportional and the stratified weighted. Lastly, the standard errors of the phi coefficients varied by method (most pairwise comparisons were significant) with the stratified weighted approach having the greater precision.

Table 4: Mean Phi Coefficients of Samples

	Random Selection	*Unadjusted Clustered*	*Adjusted Clustered*	*Stratified Proportional*	*Stratified Unweighted*	*Stratified Weighted*
Phi - Factor 1	0.994	0.981	0.983	0.995	0.985	0.995
Phi - Factor 2	0.990	0.971	0.973	0.990	0.977	0.991
Phi - Factor 3	0.982	0.955	0.954	0.982	0.979	0.984
Phi - Factor 4	0.966	0.919	0.912	0.967	0.929	0.968

Phi - Factor 5	0.942	0.855	0.852	0.943	0.891	0.951
Phi - Factor 6	0.931	0.817	0.778	0.931	0.920	0.945

Table 5: Standard Errors of Phi

	Random Selection	*Unadjusted Clustered*	*Adjusted Clustered*	*Stratified Proportional*	*Stratified Unweighted*	*Stratified Weighted*
Phi - Factor 1	0.005	0.017	0.014	0.005	0.012	0.005
Phi - Factor 2	0.007	0.023	0.020	0.007	0.014	0.007
Phi - Factor 3	0.012	0.027	0.028	0.012	0.012	0.010
Phi - Factor 4	0.028	0.057	0.061	0.027	0.049	0.025
Phi - Factor 5	0.048	0.092	0.090	0.048	0.066	0.040
Phi - Factor 6	0.055	0.126	0.153	0.055	0.046	0.041

4.2 The Results – Phi Coefficients

Most methods exhibited a positive bias (i.e. the average values were significantly greater than those in the population PCA) for the measures of magnitude of the eigenvalues. An exception for the first eigenvalue was the Stratified Unweighted method which exhibited a negative bias. The number of eigenvalues greater than one exhibited a negative bias (i.e. fewer eigenvalues were greater than one over 10,000 samples) for four of the six methods. The smallest biases were exhibited by the Stratified Weighted approach. The Stratified Weighted approach had the smaller absolute deviations for the three measures involving sums of the eigenvalues. The clustered approaches had the largest biases and the largest mean absolute deviations. The adjustment made the value of the first eigenvalue on the average larger than the population.

Table 6: Mean Eigenvalue Patterns

	Random Selection	*Unadjusted Clustered*	*Adjusted Clustered*	*Stratified Proportional*	*Stratified Unweighted*	*Stratified Weighted*
First Eigenvalue	6.273	6.463	6.621	6.273	6.049	6.260
First six	17.953	18.125	18.114	17.949	18.036	17.946
Sum of Eigenvalues >1	17.949	18.133	18.008	17.946	18.036	17.944
Number >1	5.997	6.006	5.888	5.996	6.001	5.998

Table 7: Eigenvalue Patterns Bias

	Random Selection	*Unadjusted Clustered*	*Adjusted Clustered*	*Stratified Proportional*	*Stratified Unweighted*	*Stratified Weighted*
First Eigenvalue	0.071	0.261	0.419	0.071	-0.152	0.058
First six	0.106	0.278	0.267	0.103	0.189	0.099
Sum of Eigenvalues >1	0.102	0.286	0.161	0.099	0.189	0.097
Number >1	-0.003	0.006	-0.112	-0.004	0.001	-0.002

Table 7: Eigenvalue Patterns Absolute Deviation Means

	Random Selection	*Unadjusted Clustered*	*Adjusted Clustered*	*Stratified Proportional*	*Stratified Unweighted*	*Stratified Weighted*
First Eigenvalue	0.142	0.321	0.443	0.142	0.180	0.139

First six	0.189	0.349	0.327	0.186	0.224	0.172
Sum of Eigenvalues >1	0.195	0.401	0.393	0.193	0.225	0.175
Number >1	0.010	0.109	0.186	0.010	0.002	0.005

Table 8: Standard Error of Eigenvalue Patterns

	Random	*Unadjusted*	*Adjusted*	*Stratified*	*Stratified*	*Stratified*
	Selection	*Clustered*	*Clustered*	*Proportional*	*Unweighted*	*Weighted*
First Eigenvalue	0.164	0.295	0.305	0.163	0.155	0.164
First six	0.210	0.337	0.301	0.209	0.193	0.191
Sum of Eigenvalues >1	0.229	0.415	0.460	0.228	0.196	0.201
Number >1	0.098	0.330	0.417	0.098	0.039	0.071

6. Conclusion

The study was conducted using only one data base. With 10,000 simulations per method, even a small difference was often significant. Clustering can affect a PCA, though the effects are greater for the later factors. The use of any sampling approach may tend to overestimate the amount of the variance accounted for by the first factors. Proper stratification and weighting can reduce bias and better approximate the population PCA.

The effects of clustering are more problematic. The reason is that for this data base, the intra-class correlation with respect to the clusters is rather high, and will be among the factors determining the population PCA. Efforts to control for the effect of clusters will distort the correlation between the variables. But without any control, the component pattern will be different for different samples.

An adjustment method which may work well for the first component may deteriorate faster for later components. Given that the first component seems to be reproduced well even for the unadjusted cluster sample, if one has to conduct a PCA on a cluster sample, it may be best to treat it like any other PCA, but to be guarded about the interpretation of later components.

Finally, the same can be said of PCA results as can be said of point estimates: stratification is good, but clustering is not.

References

Skinner, C.J., Holmes, D.J., and T.M.F. Smith (1986), "The Effect of Sample Design on Principal Component Analysis," ***Journal of the American Statistical Association*****, Vol. 81, No. 395, p.789-798.**