PROC PRINCOMP

The SAS procedure for carrying out a principle component analysis is PROC PRINCOMP. The basic syntax of the procedure is very simple:

PROC PRINCOMP <*options*>;

The procedure options will control printed output, as well as the actual computation of the PCA. Examples of printing options are:

NOPRINT - suppress any printed output
PREFIX - assign an alphabetic prefix to the PCA axis labels (defaut = PRIN).

Computational options include:

N - limits the PCA to N axes, where N <= the number of variables
STANDARD - standardizes the PCA scores so they all have variance = 1, and
COVARIANCE - forces computation of the PCA from the covariance matrix of the response variables.

This last option is important. By default, SAS computes PCA axes based on the correlation matrix of the specified variables. If the variables all have different scales, for example percentages vs kilograms vs miles, etc, the correlation matrix provides a standardized measure across all variables. Computing a PCA based on a non-standardized measure, such as the covariance matrix, may give misleading results. The COVARIANCE option should only be used when all the variables are of similar scales and magnitudes. Typically, the default option (correlation matrix) is the best approach.

A last set of options produces output SAS datasets. These are:

OUTSTAT= <name> - creates a SAS dataset with the eigenvalues and eigenvectors, and
OUT= <name> - produces a SAS dataset with computed PCA scores.

The OUTSTAT option is useful if PCA scores are required from a second independent data set. This might be used for PCA regression analyses. The OUT option is required to create a dataset for biplot graphics. These plots will be demonstrated later.

Other Statements

By default, PRINCOMP will compute a PCA using all numeric variables in the dataset. To control what variables are used in the procedure, a VAR statement is specified as:

VAR *var1 var2 var3 ... var n*;

where *var1 - var n* is a list of the response variables you are interested in.

Other statements include **FREQ** and **WEIGHT** which are used for summarized and weighted data forms, respectively. The **PARTIAL** statement computes a specialized PCA based on a selected set of the variables. As with all SAS procedures, a **BY** statement can also be used to compute the PCA separately for each BY variable. Note that the dataset should be sorted in the order of the BY statement before it can be used.

Example

The SAS code for a PCA on the flour viscosity data would be:

```
PROC PRINCOMP OUT = PRINS;
     VAR PEAK_VISC TROUGH_VISC FINAL_VISC BREAKDOWN
          TOTAL_SETBACK TIMEPEAK_VISC;
```

Here the OUT statement is used to create a dataset, PRINS, which contains the computed PCA scores. These will be used later to produce a biplot of the PCA axes. The printed output from this program is listed below. The first section reports the number of observations and variables used along with the simple summary stats (mean and standard deviation) for each variable.

```
The PRINCOMP Procedure

Observations         225
Variables              6


Simple Statistics

            Peak_Visc       Trough_Visc        Final_Visc

Mean       553.8751862       267.3611120       359.1992600
StD        242.2611888       143.1644188       131.9776057

Simple Statistics

                                  Total_         TimePeak_
            Breakdown           Setback              Visc

Mean       286.4740742       91.79814933       5.397556889
StD        196.6361784       53.98203963       0.965913559
```

The second section of the printout gives the Pearson correlation matrix for the six variables. This is the correlation matrix that is used by the default program to compute the PCA.

Correlation Matrix

	Peak_Visc	Trough_Visc	Final_Visc	Breakdown	Total_Setback	TimePeak_Visc
Peak_Visc	1.0000	0.5842	0.8002	0.8065	0.4061	-.3004
Trough_Visc	0.5842	1.0000	0.9263	-.0088	-.3890	-.1791
Final_Visc	0.8002	0.9263	1.0000	0.3110	-.0133	-.2734
Breakdown	0.8065	-.0088	0.3110	1.0000	0.7837	-.2399
Total_Setback	0.4061	-.3890	-.0133	0.7837	1.0000	-.1944
TimePeak_Visc	-.3004	-.1791	-.2734	-.2399	-.1944	1.0000

The last section of the output provides the eigenvalues and eigenvectors (loadings) for each axis. In this case the cumulative percent of variability accounted for by the first two axes is 82.7%. This indicates that the first two axes describe the majority of the variability among these six variables. Therefore, we "retain" or interpret only PCA axes 1 and 2. Another method for determining the number of axes to retain is to use only those axes with eigenvalues greater than or equal to 1.0.

The eigenvectors indicate the relative importance of each variable within the individual axes. Importance is determined based on the absolute magnitude of the eigenvector coefficients or loadings. Deciding which loadings are "large" is a subjective decision and there are no rules for picking out coefficients. For example, PCA axis1 appears to have large loadings on the Peak Viscosity, Final Viscosity, and Breakdown variables. Thus, axis1, which accounts for 50% of the variability, appears to be associated with the RVA high points and the initial drop in viscosity. The second axis has large loadings on the variables Trough Viscosity and Total Setback which are associated with the RVA low point and recovery phase in RVU measurements. Note that the loading values for these variables are also opposite in sign. This reflects the negative correlation between Trough Viscosity and Total Setback.

	Eigenvalue	Difference	Proportion	Cumulative
1	**3.00088240**	1.03919934	**0.5001**	0.5001
2	**1.96168306**	1.10533355	**0.3269**	**0.8271**
3	0.85634951	0.67527380	0.1427	0.9698
4	0.18107571	0.18106638	0.0302	1.0000
5	0.00000933	0.00000933	0.0000	1.0000
6	0.00000000		0.0000	1.0000

Eigenvectors

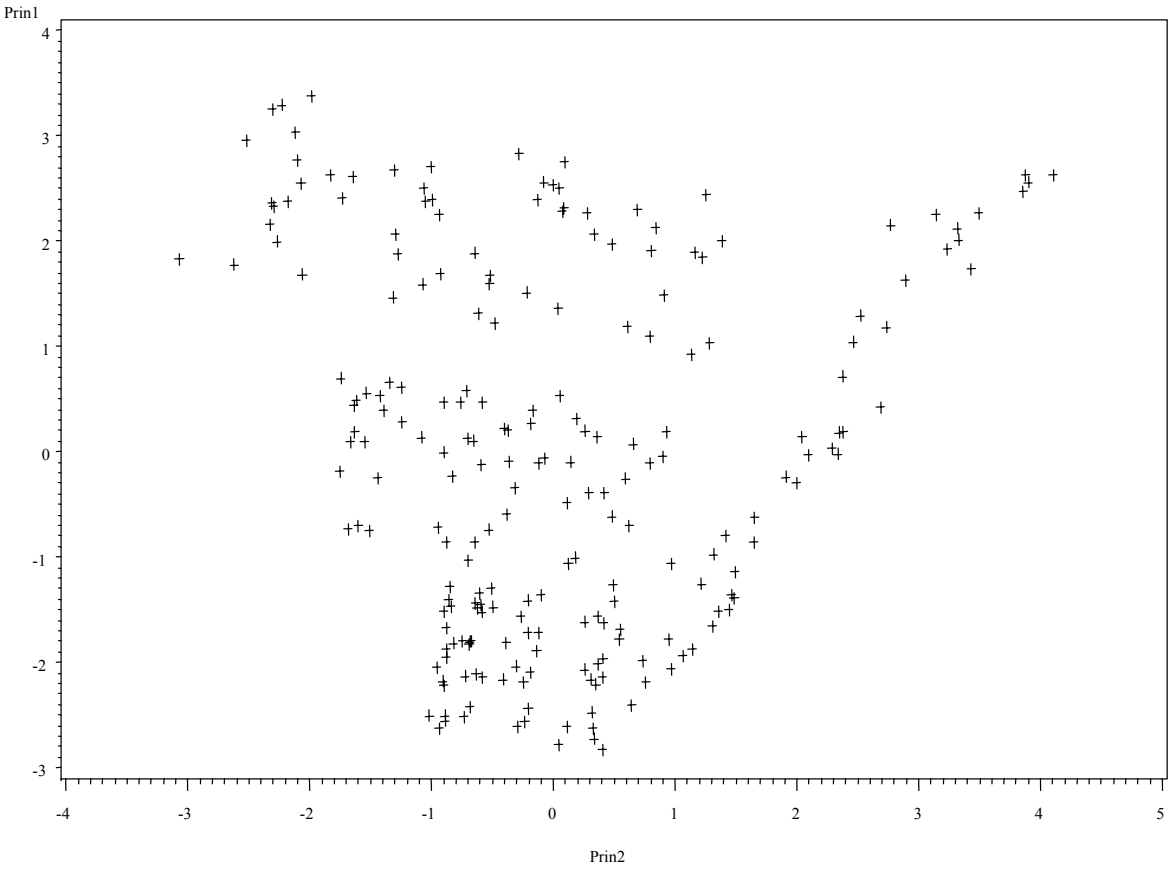
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6
Peak_Visc	**0.562975**	0.042138	0.175696	-.323760	-.216566	-.706182
Trough_Visc	0.370800	**-.544618**	0.061956	0.111615	0.741351	-.000000
Final_Visc	**0.496443**	-.337964	0.087558	0.405556	-.564960	0.384710
Breakdown	**0.423512**	0.448707	0.170984	-.480218	0.175817	0.573187
Total_Setback	0.229896	**0.619092**	0.048406	0.695243	0.231098	-.157356
TimePeak_Visc	-.258643	-.052747	0.962313	0.065378	0.000350	-.000000

One of the more useful aspects of PCA is the biplot analysis. These plots can be produced with the SAS/GRAPH code as follows:

```
SYMBOL1 I = NONE V=PLUS C = BLACK;
SYMBOL2 I = NONE V=SQUARE C = BLUE;
SYMBOL3 I = NONE V=CIRCLE C = GREEN;
SYMBOL4 I = NONE V=STAR C = RED;
SYMBOL5 I = NONE V=TRIANGLE C = PURPLE;

PROC GPLOT DATA=PRINS;
     PLOT PRIN1*PRIN2 = 1;
```

The symbol statements in this code simply define 5 different plotting symbols. The PROC GPLOT statements plot the first two PCA axes (PRIN1 and PRIN2) of the output dataset PRINS, against one another. The resulting plot is given below.

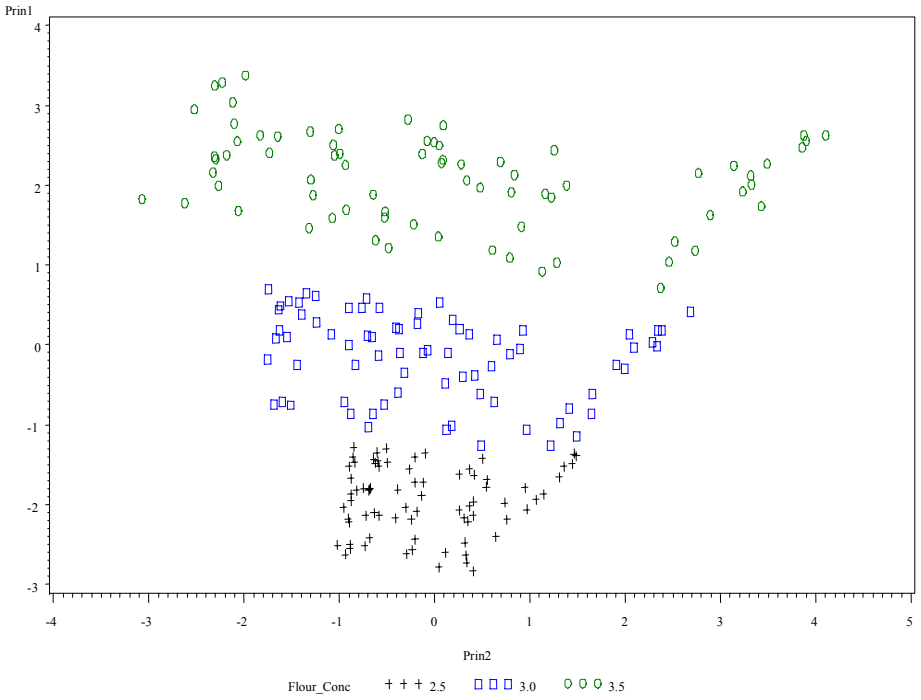


This plot indicates there may be some grouping or structure present in the axes. One way to investigate this type of structure is to color code the plot based on the values of another variable. For example, the PROC GPLOT code could be modified to:

```
PLOT PRIN1*PRIN2 = FLOUR_CONC;
```

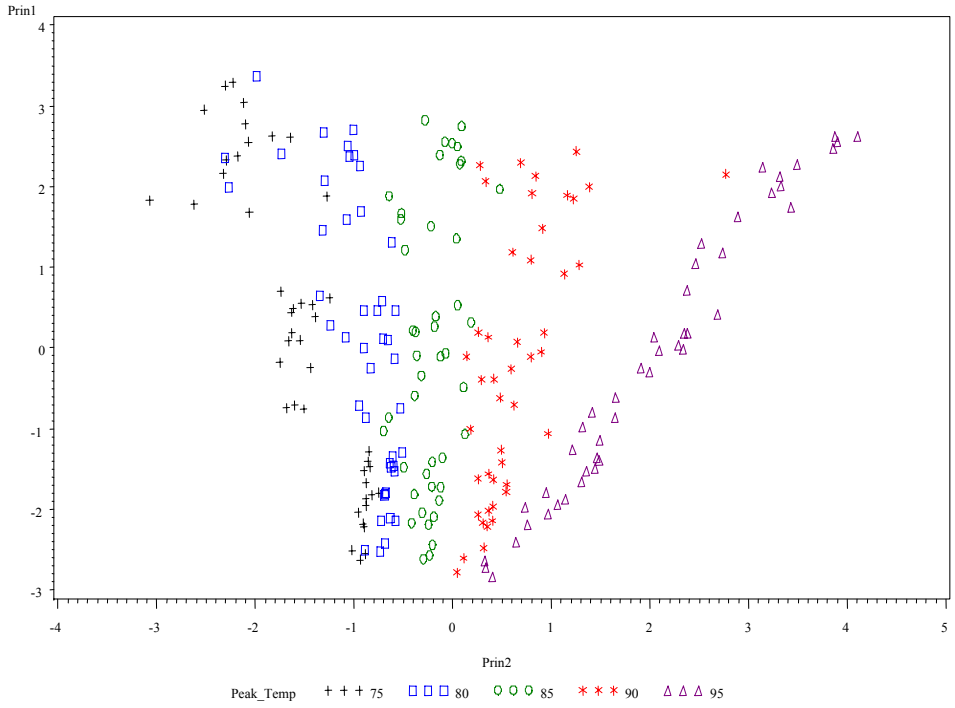
where the data points will now be coded according to Flour Concentration levels. This plot is

more informative:



We can now see three broad bands extending across the plot. Note that the bands change relative to PCA axis1. Thus, we might conclude Flour Concentration is affecting Peak Viscosity, Final Viscosity and Breakdown. A similar plot for Peak Temperature can be made using:

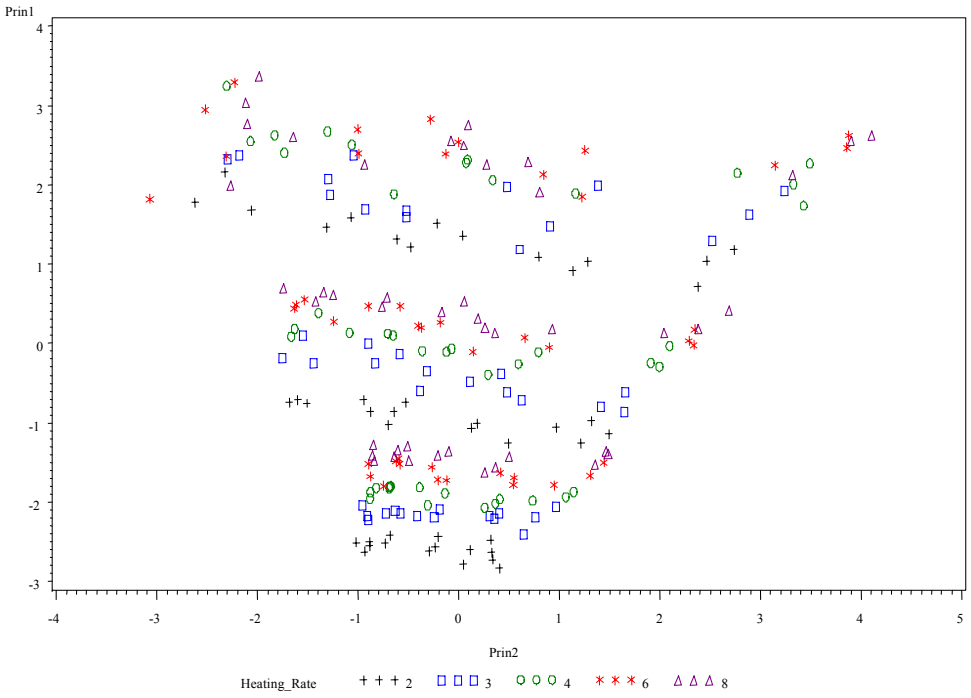
```
PLOT PRIN1*PRIN2 = PEAK_TEMP;
```



Here, the banding is vertical and changes relative to the level of PCA axis2. Peak Temperature, therefore, is affecting the Trough Viscosity and Total Setback. We can also note that the "slope" of the bands changes from a negative direction at a Peak Temperature of 75* C to a positive one at 95*C. This indicates that the relationship between axis1 and axis 2 variables changes depending on the setting of Peak Temperature.

Finally, a fourth plot can be made with:

```
PLOT PRIN1*PRIN2 = HEATING_RATE;
```



In this case, small horizontal bands are evident which also appear to repeat within the Flour Concentration bands. This may indicate a potential interaction between Flour Concentration and Heating Rate.

Typically, a PCA analysis would not be conducted on a designed experiment such as the one described above. However, PCA was used in this example because the structure of the data, i.e. the experimental factors, were known prior to analysis. Imagine that the structure was unknown, as might be the case with a survey. In that situation, these groupings and patterns would provide valuable insight into the data structure.

PROC FACTOR

Factor analysis is very similar to PCA. With factor analysis, however, axis placement is less restrictive. The axes may only approximate the longest dimensions of the data (orthogonal rotations or may not be required to be perpendicular to each other (oblique rotations). The decision of which rotation to use is left to the user and is usually aimed at maximizing the interpretation or meaning of each axis. Thus, factor analysis is very subjective and, from a objective statistical standpoint, not very attractive. For completeness the SAS code for factor analysis will be provided here, but users should exercise caution and seek professional consultation before usage.

The SAS code for factor analysis is similar to the code above for PCA:

```
PROC FACTOR <options>;
    VAR var1 var2 var 3 ... var n;
```

The options here are much more numerous than PRINCOMP, but the basics are the same. For example:

NOPRINT
COVARIANCE
OUTSTAT, and
OUT.

perform the same functions as in PRINCOMP. Other options are available and allow specification of the axis rotation type:

ROTATE =
PREROTATE =

It should be noted that a ROTATE=none option is equivalent to PCA and will give the same results as PROC PRINCOMP. This is the default for PROC FACTOR.

The remaining options relate to the details of rotation types, estimation procedures, and plotting/printing output. If factor analysis is required for a research project, users should seek the advice of a statistician before conducting the analysis.