Introduction:

The purpose of the paper is to break the traditional paradigm of doing market basket analysis (MBA) using association rules and to show that we can also use factorial algorithms like PROC VARCLUS to arrive market baskets. This paper shows step by step ways of doing MBA using VARCLUS method. While doing so this paper also throws insight on the usage of PROC VARCLUS procedure in arriving product groups that are bought together more frequently by the customers.

Background:

Market Basket Analysis is a data mining tool to identify the co-occurrence or co-existence of nominal or categorical observations. It is been widely used to identify purchasing pattern of customers in a retail stores using transaction data. The outcome of the analysis will give rules like “People who bought items on set X are often also bought items on set Y”.

The most widely used techniques to conduct market basket Analysis is association rules technique. Although it is most widely used techniques, the following are the shortcoming of the application of association rules to arrive market baskets. They are:

1. Both independent and dependent data are nominal or categorical type. Therefore, what we can do is only count the frequency of pattern. Such as market basket analysis using association rule is sometimes called as mining frequent pattern. If the data are quantitative, we need to categorize it into some interval (but the meaning is actually nominal) such as: age 0 to 1, age 1-5, age 5 to 12, age 12 to 19, etc.
2. In Market Basket Analysis, we usually do not consider about the number of each items that the customers bought. Whether a customer buys one kg of apple or 10 kg of apple would be considered as the same set of apple.
3. We do not use all transactions that are recorded. Only transactions of purchase of more than one item are considered as data. Transactions of single item are not used for the analysis.
4. The input data are assumed to be clean from error and noise.

1 http://people.revoledu.com/kardi/tutorial/MarketBasket/Characteristics.htm
In this paper we are trying to demonstrate, how to do MBA using Factorial algorithm (i.e., PROC VARCLUS – SAS Procedure). Using VARCLUS also helps to overcome the shortcomings of step 1 to 3 mentioned above. In addition to that we also found that using factor analysis technique helps in saving lot of time in terms of computing resources like space and server run time.

**Data Preparation**

**Data Format**

The input SAS data must contain three columns of variables. The first required column is customer identifier. This is typically unique to the Merchandise Division (MD) the customer made purchase. The second required column is MD identifier. This is typically a product or product grouping identifier. The third required column is purchase amount made on each of the MDs by the customer. The data set should contain one row of data for each unique combination of customer identifier and MD identifier with respect to the purchase amount. A visual example of the base data can be seen in Figure -1.

<table>
<thead>
<tr>
<th>CUSTOMER IDENTIFIER</th>
<th>PRODUCT GROUP IDENTIFIER</th>
<th>PURCHASE AMOUNT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customer 1</td>
<td>Product Group A</td>
<td>$ XXX</td>
</tr>
<tr>
<td>Customer 2</td>
<td>Product Group B</td>
<td>$ XXX</td>
</tr>
<tr>
<td>Customer 2</td>
<td>Product Group C</td>
<td>$ XXX</td>
</tr>
<tr>
<td>Customer 3</td>
<td>Product Group A</td>
<td>$ XXX</td>
</tr>
<tr>
<td>Customer 3</td>
<td>Product Group B</td>
<td>$ XXX</td>
</tr>
<tr>
<td>Customer 3</td>
<td>Product Group D</td>
<td>$ XXX</td>
</tr>
</tbody>
</table>

Once the base dataset has been created using the transactional data using the methodology described above, one need to prepare the data for running the PROC VARCLUS procedure. The following program would help us to do that.

```sas
PROC SUMMARY DATA=<Base Dataset> NWAY MISSING;
CLASS <customer identifier product identifier>>;
VAR <Purchase Amount>;
OUTPUT OUT=< dataset-1> (drop = _type_ _freq_)SUM=;
RUN;

PROC TRANSPOSE DATA=<dataset-1> OUT=<dataset-2>;
ID <Product_Group_Identifier>;
VAR <Purchase Amount>;
BY <Customer Identifier>;
RUN;
```
The input dataset for the PROC VARCLUS procedure will transpose the data to have each product group identifier as one column. At this level we would not be having any duplicate customer identifier in the dataset. Such as, each row is unique to the customer identifier. A visual example of the base data can be seen in Figure -2.

<table>
<thead>
<tr>
<th>CUSTOMER IDENTIFIER</th>
<th>PRODUCT GROUP - A PURCHASE AMOUNT</th>
<th>PRODUCT GROUP - B PURCHASE AMOUNT</th>
<th>PRODUCT GROUP - C PURCHASE AMOUNT</th>
<th>PRODUCT GROUP - D PURCHASE AMOUNT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customer 1</td>
<td>$ XXX</td>
<td>$ XXX</td>
<td>$ XXX</td>
<td>$ XXX</td>
</tr>
<tr>
<td>Customer 2</td>
<td>$ XXX</td>
<td>$ XXX</td>
<td>$ XXX</td>
<td>$ XXX</td>
</tr>
<tr>
<td>Customer 3</td>
<td>$ XXX</td>
<td>$ XXX</td>
<td>$ XXX</td>
<td>$ XXX</td>
</tr>
</tbody>
</table>

Once the base dataset has been transformed using PROC transpose, we need clean the data by missing value treatment and outlier removal / treatment. Missing value treatment implies replacing the null values into 0. This is because if a customer did not make any purchase in a particular MD can also be said as the customer made 0 purchases in that MD. Outlier analysis implies removing extraordinary purchase amount or replacing the extraordinary purchase amount by the average purchase amount in that particular MD by all the customers.

Upon completing the data cleaning is done we can run the PROC VARCLUS procedure. The following program would help us to do that.

```
PROC VARCLUS DATA=final_data2 OUTTREE=tree;
VAR sales:;
RUN;
```

**PROC VARCLUS to Arrive MD Groups:**

The VARCLUS procedure attempts to divide a set of variables into nonoverlapping clusters in such a way that each cluster can be interpreted as essentially unidimensional. For each cluster, PROC VARCLUS computes a component that can be either the first principal component or the centroid component and tries to maximize the sum across clusters of the variation accounted for by the cluster components. A large set of variables can often be replaced by the set of cluster components with little loss of information.

PROC VARCLUS creates an output data set that can be used with the SCORE procedure to compute component scores for each cluster. A second output data set can be used by the TREE procedure to draw a tree diagram of hierarchical clusters².

---

² [http://www.okstate.edu/sas/v8/saspdf/stat/chap68.pdf](http://www.okstate.edu/sas/v8/saspdf/stat/chap68.pdf)
For our exercise, PROC VARCLUS attempts to divide the set of MD variables into nonoverlapping MD groups. This can be interpreted as the each unidimensional cluster contains MDs that are highly correlated among themselves but have least correlated with other unidimensional clusters. In other way, each MD variable cluster contains the MDs are most likely bought together as compared to the MDs in other MD variable clusters.

**Methodology Summary:**
In summary the methodology of doing MBA analysis using factorial algorithm like PROC VARCLUS would look like this.

1. Prepare customer level MD purchase history data for the customers who were purchase active for the last 12 month time period.
2. Transform the data by product group. So that we will have MD purchase data as columns and customer level data as rows.
3. Do the data cleaning such as missing value treatment, outlier analysis of the MD variables.
4. Run PROC VARCLUS with the MD level purchase history data.
5. Form Merchandise Division “Groups” based on PROC VARCLUS factor loading and business logic of grouping Merchandise Divisions.
6. Validate the stability of “Clusters” by developing the MD Groups based on MD level purchase history data at quarterly level or do out of time period validation.

**Conclusion:**
In this paper we tried to break the industry practice of arriving market basket analysis using association rules by using the factorial algorithm such as PROC VARCLUS – SAS procedure, which is a principle component factor analysis method.

**References:**
1. Kardi Teknomo, PhD. “Characteristics of Market Basket Analysis”,
2. Chapter-68, The VARCLUS Procedure, SAS user guide,