Paper 320-2008

TWO-STAGE VARIABLE CLUSTERING FOR LARGE DATA SETS

Taiyeong Lee, David Duling, Song Liu, and Dominique Latour
SAS Institute Inc., Cary, NC

ABSTRACT

In data mining, principal component analysis is a popular dimension reduction technique. It also provides a good remedy for the multicollinearity problem, but its interpretation of input space is not as good. To overcome the interpretation problem, principal components (cluster components) are obtained through variable clustering, which was implemented with PROC VARCLUS. The procedure uses oblique principal components analysis and binary iterative splits for variable clustering, and it provides non-orthogonal principal components. Even if this procedure sacrifices the orthogonal property among principal components, it provides good interpretable principal components and well-explained cluster structures of variables. However, the PROC VARCLUS implementation is inefficient to deal with high-dimensional data. We introduce the two-stage, variable clustering technique for large data sets. This technique uses global clusters, sub-clusters, and their principal components.

INTRODUCTION

Dimension reduction is one of most important data mining tasks to handle data sets with a very large number of variables. Some easy and common, supervised dimension reduction tasks can be achieved through simple linear regression, that is, by using R-squares between dependent and independent variables, stepwise regression, and other variants of the regression method. The methods are also used as preprocessing methods of some noble dimension techniques when the number of variables is extremely large. Another popular method is an unsupervised technique that uses principal components analysis. This technique gives very successful dimension reduction results and remedies the multicollinearity problem. However it suffers from its interpretation for input space and some computation problems in the eigenvalue calculation when the dimension of input space is very large. To overcome those difficulties, we can use a method that combines supervised and non-supervised methods, for example, a simple variable selection that uses an R-square or a Chi-Square test with its target variable, then another dimension reduction technique such as principal components analysis.

Dimension reduction can be achieved through variable clustering, too. There are two typical types of variable clustering techniques. One method is to apply common clustering techniques to any distance matrix of the variables. This method is usually

applied to observation clustering. The other method is using variable structure from factor analysis or principal components analysis. The former is widely used and very intuitive. Its performance depends largely on the type of clustering algorithm that is used. The latter is a more expensive process than the former because it requires eigenvalue decomposition and a certain iterative optimization process such as factor rotation. In our experience, the variable structure method gives better results more often than the clustering techniques method.

One of the latter methods is used with PROC VARCLUS in SAS/STAT® software. This method provides an iterative variable clustering that uses oblique rotation of principal components. The procedure generates variable cluster structures, identifies key variables within each cluster, and provides non-orthogonal principal components that are called cluster components. The cluster components give much better interpretation than regular principal components because they consist of only the variables in each cluster. PROC VARCLUS has been successfully used by Cerrito *et. al.* (1999), Sanche and Lonergan (2006), and Nelson (2001).

D'Agostino and Russell (2005) also show several variants of variable-clustering methods using latent variables that include the PROC VARCLUS algorithm. However, not all of the papers addressed large data sets for variable clustering, and no benchmarking for large data sets was reported. We did a benchmarking based on the PROC VARCLUS algorithm, and found that it is not scalable at all. The report is shown in a section of this paper. Because the procedure requires a lot of calculation resources for very large data sets with, for example, more than 3,000 variables, it is inappropriate for data mining tasks. In order to overcome the problem, we therefore propose the two-stage variable clustering, which combines two types of variable clustering methods that are described in the next section.

VARIABLE CLUSTERING

As we mentioned in the introduction, two methods of clustering variables are widely used: one is clustering based on a distance matrix, and the other is using latent variables. Both methods are summarized below.

Method 1: Variable clustering that is based on a distance matrix

1. Calculate any distance matrix of the variables (for example, correlation matrix).
2. Apply any (observational) clustering algorithm to the distance matrix.
3. Obtain clusters that contain homogeneous variables.
4. (Optional) Calculate cluster components (or first principal components) from each cluster.

Method 2: Variable clustering that is based on latent variables

1. Start with all variables to find the first two principal components.
2. Perform an orthoblique rotation (quartimax rotation) on eigenvectors.
3. Assign each variable to the rotated component with which it has the higher squared correlation.
4. Variables are iteratively reassigned to clusters to try to maximize the variance accounted for by the cluster components.
5. Stop the iterative reassignment when some criteria are met.

Some variants of variable clustering that use latent variables are found in the section of "Cluster Analysis, Variables" in *Encyclopedia of Biostatistics* (1998). However, because we are interested in PROC VARCLUS, Method 2 shows the method of PROC VARCUS by quoting the algorithm description from the SAS/STAT® document. More details can be found in the PROC VARCLUS section of the SAS/STAT® document. The *single-stage method* in this paper refers to Method 2.

First we compare both methods with the small data set generated by the following SAS code. The data set contains two groups of variables. There exists apparent collinearity among the first three variables for each group. The pair-wise correlations among them are exactly 1 because we didn't add noise variables. The other two variables have a nonlinear relationship with the first three variables. The initial variables, x1 and x6, are independently generated for each group.

```
data example1;
   do i = 1 to 100;
          x1=rannor(12334);  x2=0.9*x1; x3=0.7*x1; x4=x1*x1; x5=x1*x1*x1;
          x6=rannor(56757);  x7=0.9*x6; x8=0.7*x6; x9=x6*x6; x10=x6*x6*x6;
         output;
   end;
run;
```

Using the example data set, the hierarchical clustering of PROC CLUSTER that is based on a correlation matrix produces the following dendrogram of variable clusters. When we cut two clusters from the dendrogram, the variable X4 is classified into the cluster that X1 does not belong to even if X4 is a function of x1 rather than X6. When three clusters are chosen, x4 and x9 are grouped into one cluster, even though they are functions of different independent random variables.

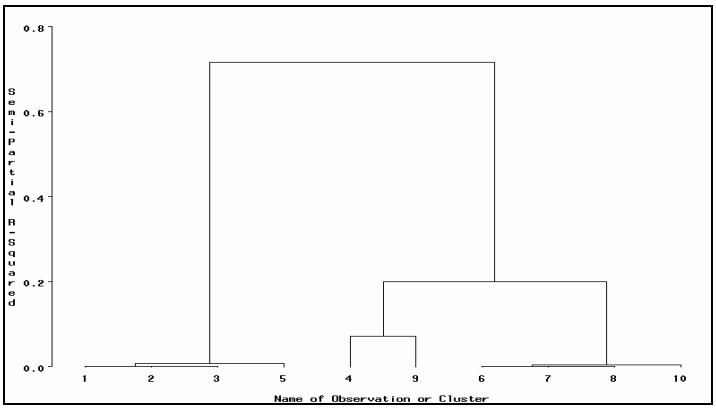


Figure 1. Dendrogram from Variable Clustering Using a Distance Matrix

It seems that the simple method could not satisfactorily reveal the variable structure in this example. In comparison, the latent variable method shows clearly separated clusters. The cluster analysis using latent variables is explained through the variable clustering node in SAS® Enterprise Miner™, which implements PROC VARCLUS. The default tool setting is used with the same EXAMPLE1 data set.

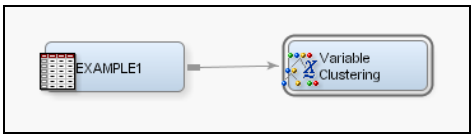


Figure 2. Single-Stage Variable Clustering

The single-stage variable clustering starts with all variables, and the first two principal components are extracted. The factors are rotated by quartimax method, and squared correlation between variables and the rotated components are calculated. The following cluster structure shows the correlations.

Cluster Structure

Cluster	1	2
x1	-.034619	0.977479
x2	-.034619	0.977479
x3	-.034619	0.977479
x4	0.106374	0.461054
x5	-.069813	0.902560
x6	0.985842	-.026885
x7	0.985842	-.026885
x8	0.985842	-.026885
x9	0.325550	-.064554
x10	0.932360	-.015957

Figure 3. Calculated Correlations between Variables

Look at how the node made the clusters through the cluster structure. For the variables from X1 to X5, the correlations with the first component are greater than those with the second component. So the variables from X1 to X5 are grouped into one cluster. For the variables of x6 to x10, the correlations with the second components are greater than

those with the first component. Consequently, they are grouped into another cluster. Only one binary split is required to meet the default stopping criterion because all second eigenvalues are less than 1.

Cluster Summary for 2 Clusters

Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained	Second Eigenvalue
1	5	5	3.890931	0.7782	0.9504
2	5	5	3.89358	0.7787	0.8937

Total variation explained = 7.784512 Proportion = 0.7785

Figure 4. Cluster Summary for Two Clusters

Therefore two clusters are obtained and the corresponding cluster plot is shown in Figure 5.

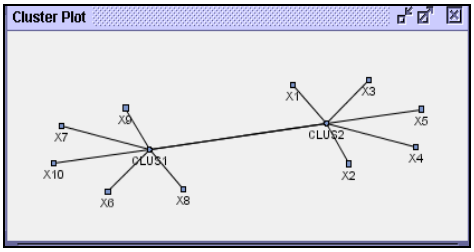


Figure 5. Cluster Plot by Single-Stage Method

Within each cluster, the first principal component is recalculated as its cluster component. The standardized scoring coefficients of variables show exactly zero in the clusters except for its own cluster, which increases interpretability of cluster components in dimension reduction.

Standardized Scoring Coefficients

Cluster	1	2
x1	0.000000	0.251049
x2	0.000000	0.251049
x3	0.000000	0.251049
x4	0.000000	0.118414
x5	0.000000	0.231807
x6	0.253369	0.000000
x7	0.253369	0.000000
x8	0.253369	0.000000
x9	0.083669	0.000000
x10	0.239624	0.000000

Figure 6. Standardized Scoring Coefficients

If we change the stopping criterion from the default to 0.9 in proportion to the variation that we explained, the change results in two more splits, so four clusters are obtained.

Cluster Summary for 1 Cluster

Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained	Second Eigenvalue
1	10	10	4.019585	0.4020	3.7890

Cluster 1 will be split because it has the smallest proportion of variation explained, 0.401959, which is less than the PROPORTION=0.9 value.

Cluster Summary for 2 Clusters

Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained	Second Eigenvalue
1	5	5	3.890931	0.7782	0.9504
2	5	5	3.89358	0.7787	0.8937

Cluster 1 will be split because it has the smallest proportion of variation explained, 0.778186, which is than the PROPORTION=0.9 value.

Cluster Summary for 3 Clusters

Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained	Second Eigenvalue
1	4	4	3.810161	0.9525	0.1898
2	5	5	3.89358	0.7787	0.8937
3	1	1	1	1.0000	

Cluster 2 will be split because it has the smallest proportion of variation explained, 0.778716, which is less than the PROPORTION=0.9 value.

Cluster Summary for 4 Clusters

Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained	Second Eigenvalue
1	4	4	3.810161	0.9525	0.1898
2	4	4	3.727078	0.9318	0.2729
3	1	1	1	1.0000	
4	1	1	1	1.0000	

No cluster meets the criterion for splitting.

Figure 7. Cluster Summaries

This result is contrary to the cluster result by the method 1. Even if we split the variables more than two clusters, the x4 and x9 will not be grouped into the same cluster. They will stay alone as CLUS4 and CLUS3, but the clusters are close to the cluster whose source random variable is. The following cluster plot shows a clear view of clusters.

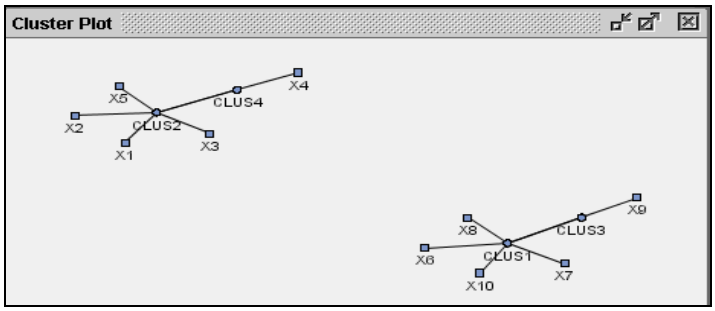


Figure 8. Cluster Plot with New Stopping Criteria

From this particular example, we see that PROC VARCLUS produces more accurate cluster results than the distance-based method. It does not mean, however, that the PROC VARCLUS is always superior to the other method because no rigorous comparison has been reported.

TWO-STAGE VARIABLE CLUSTERING

As we mentioned in the introduction, PROC VARCLUS is not efficient with very large data sets. So we propose a combined method that combines the two methods that were described in the previous section for such a large data set. First we apply the simple method to get global clusters, and then apply the latent variable method to get sub-

clusters. Finally we create a single tree of global clusters and sub-clusters. The combined method is described in two stages below.

STAGE 1: Variable clustering based on a distance matrix

1. Calculate the correlation matrix of the variables.
2. Apply a hierarchical clustering algorithm to the correlation matrix.
3. Using a predefined cluster number, cluster variables into homogeneous groups. The cluster number is generally no more than the integer value of (nvar/100+2). These clusters are called global clusters.

STAGE 2: Variable clustering based on latent variables

1. Run PROC VARCLUS with all variables within each global cluster as you would run a single-stage, variable clustering task.
2. For each global cluster, calculate the global cluster components, which are the first principal component of the variables in its cluster.
3. Create a global cluster structure using the global cluster components and the same method as 1 at STAGE 2.
4. Form a single tree of variable clusters from 1 and 3.

To show the performance of two-stage variable clustering with a real data set, we used the ISOLET (Isolated Letter Speech Recognition) data from the UCI Machine Learning Repository. According to the data description, 150 subjects spoke the name of each letter of the alphabet twice, so 52 training examples are generated from each speaker. The speakers are grouped into sets of 30 each. The data set was divided into a training set (120 subjects) and a test set (30 subjects). Three examples are missing. Numbers of instances are 6238 and 1559 for the training set and the test set, respectively. The number of attributes is 617 plus 1 for target. We are interested in the unsupervised variable clustering rather than the classification task.

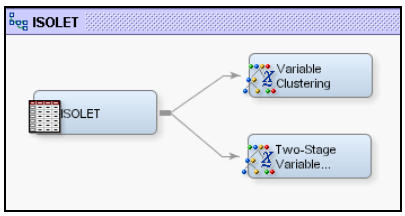


Figure 9. Variable Clustering with the ISOLET Data

First we suppress the two-stage option and run the variable clustering node. A partial result shows that Cluster1 among 108 clusters contains 13 variables and its cluster component. The total number of cluster components is 108, which means the dimension reduction rate is about 1/6.

Variable Selection Table

Cluster	Variable	R-Square With Own Cluster Component	Next Closest Cluster	R-Sqaure with Next Cluster Component	Type	Label	1-R2 Ratio	Variable Selected
CLUS1	CLUS1	1	CLUS11	0.853073	ClusterComp	Cluster 1	0	YES
CLUS1	F71	0.944577	CLUS11	0.73683	Variable		0.210598	NO
CLUS1	F72	0.927649	CLUS11	0.760116	Variable		0.301608	NO
CLUS1	F103	0.91753	CLUS11	0.797913	Variable		0.408089	NO
CLUS1	F73	0.90674	CLUS11	0.789755	Variable		0.44358	NO
CLUS1	F39	0.882728	CLUS27	0.761691	Variable		0.492102	NO
CLUS1	F70	0.887957	CLUS101	0.794358	Variable		0.544847	NO
CLUS1	F294	0.843883	CLUS88	0.731895	Variable		0.582299	NO
CLUS1	F295	0.848236	CLUS11	0.774751	Variable		0.673761	NO
CLUS1	F104	0.903503	CLUS11	0.858146	Variable		0.680257	NO
CLUS1	F74	0.872712	CLUS11	0.815227	Variable		0.688885	NO
CLUS1	F102	0.866425	CLUS88	0.810111	Variable		0.703435	NO
CLUS1	F462	0.856253	CLUS47	0.811	Variable		0.760567	NO
CLUS1	F75	0.779569	CLUS1	0.779569	Variable		1	NO

Figure 10. Variable Selection Table

The result from two-stage variable clustering shows seven global clusters, and the total number of sub-clusters is 105. There are105 cluster components to be used. We now pick up variable F71 from the CLUS 1 of the result of single-stage method and look for the same variable in the results of the two-stage method. The variable is found at the global cluster 1 and sub-cluster 9 (GC1_CLUS9). Because GC1_CLUS9 has only eight variables, we need one more sub-cluster for the comparison of clustered variables. The table in Figure 12 shows the next closest sub-cluster, which is GC1_CLUS2.

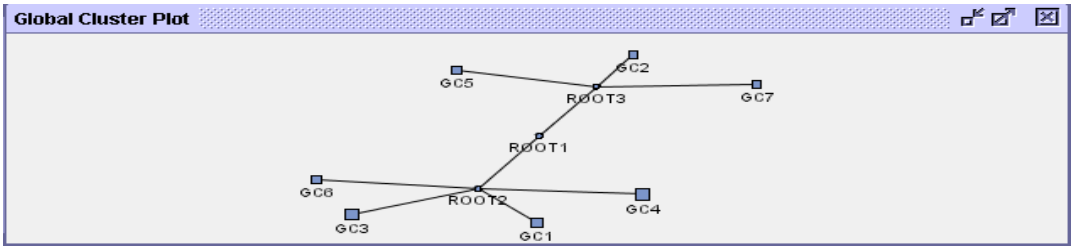


Figure 11. Cluster Plot from Two-Stage Clustering

GClsuter:1 Variable Selection

Global Cluster	Cluster	Variable	R-Square With Own Cluster Component	Next Closest Cluster	R-Sqaure with Next Cluster Component
GC1	GC1_CLUS9	F39	0.957466	GC1_CLUS2	0.781025
GC1	GC1_CLUS9	F40	0.925668	GC1_CLUS4	0.797466
GC1	GC1_CLUS9	F71	0.947602	GC1_CLUS2	0.864955
GC1	GC1_CLUS9	F72	0.937198	GC1_CLUS2	0.864188
GC1	GC1_CLUS9	F38	0.902538	GC1_CLUS10	0.834408
GC1	GC1_CLUS9	F70	0.883329	GC1_CLUS2	0.805279
GC1	GC1_CLUS9	F73	0.898559	GC1_CLUS2	0.85567
GC1	GC1_CLUS9	F7	0.862762	GC1_CLUS4	0.842419
GC1	GC1_CLUS9	GC1_CLUS9	1	GC1_CLUS2	0.867088

Figure 12. Variables in GC1_CLUS2

GCIsuter:1 Variable Selection

Global Cluster	Cluster	Variable	R-Square With Own Cluster Component
GC1	GC1_CLUS2	F104	0.930555
GC1	GC1_CLUS2	F136	0.883044
GC1	GC1_CLUS2	F103	0.922672
GC1	GC1_CLUS2	F135	0.889281
GC1	GC1_CLUS2	F105	0.900835
GC1	GC1_CLUS2	F295	0.871666
GC1	GC1_CLUS2	F296	0.822253
GC1	GC1_CLUS2	F294	0.837916
GC1	GC1_CLUS2	F102	0.850534
GC1	GC1_CLUS2	F167	0.781305
GC1	GC1_CLUS2	F462	0.84076
GC1	GC1_CLUS2	F74	0.838499
GC1	GC1_CLUS2	F75	0.762565

Figure 13. Variables in GC1_CLUS9

The following table shows that all of the 13 variables from the single-stage method are covered by the two sub-clusters from the two-stage method. In other words, the CLUS11 is a subset of the two closest sub-clusters from the two-stage method.

Table 1. Comparison of Clustered Variables between Single-Stage and Two-Stage Clustering

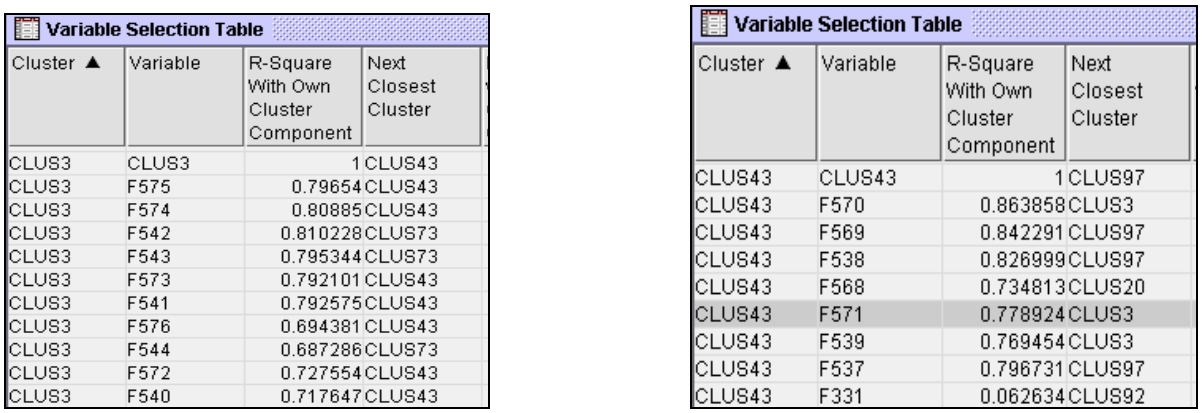
Single Stage			**F39**		**F70**	**F71**	**F72**	**F73**	**F74**	**F75**	**F102**	**F103**	**F104**					**F294**	**F295**		**F462**
Two Stage	**F7**	**F38**	**F39**	**F40**	**F70**	**F71**	**F72**	**F73**	**F74**	**F75**	**F102**	**F103**	**F104**	**F105**	**F135**	**F136**	**F167**	**F294**	**F295**	**F296**	**F462**

Here is another example. Look at a cluster from the two-stage clustering first: GCLUSTER 4 contains 11 variables. The CLUS3 from single-stage has 10 variables among them, and one variable (F571) is found at the next closest cluster (CLUS43) of CLUS3.

GCIsuter:4 Variable Selection

Global Cluster	Cluster ▲	Variable	R-Square With Own Cluster Component	Next Closest Cluster
GC4	GC4_CLUS1	F574	0.810167	GC4_CLUS20
GC4	GC4_CLUS1	F575	0.790374	GC4_CLUS18
GC4	GC4_CLUS1	F573	0.805324	GC4_CLUS20
GC4	GC4_CLUS1	F542	0.790302	GC4_CLUS18
GC4	GC4_CLUS1	F541	0.783289	GC4_CLUS18
GC4	GC4_CLUS1	F543	0.770736	GC4_CLUS18
GC4	GC4_CLUS1	F572	0.765083	GC4_CLUS20
GC4	GC4_CLUS1	F576	0.684804	GC4_CLUS14
GC4	GC4_CLUS1	F544	0.665679	GC4_CLUS18
GC4	GC4_CLUS1	F540	0.727585	GC4_CLUS20
GC4	GC4_CLUS1	F571	0.690223	GC4_CLUS20
GC4	GC4_CLUS1	GC4_CLUS1	1	GC4_CLUS20

Figure 14. GC4_CLUS1 from Two-Stage Clustering



Variable Selection Table

Cluster ▲	Variable	R-Square With Own Cluster Component	Next Closest Cluster
CLUS3	CLUS3	1	CLUS43
CLUS3	F575	0.79654	CLUS43
CLUS3	F574	0.80885	CLUS43
CLUS3	F542	0.810228	CLUS73
CLUS3	F543	0.795344	CLUS73
CLUS3	F573	0.792101	CLUS43
CLUS3	F541	0.792575	CLUS43
CLUS3	F576	0.694381	CLUS43
CLUS3	F544	0.687286	CLUS73
CLUS3	F572	0.727554	CLUS43
CLUS3	F540	0.717647	CLUS43

Variable Selection Table

Cluster ▲	Variable	R-Square With Own Cluster Component	Next Closest Cluster
CLUS43	CLUS43	1	CLUS97
CLUS43	F570	0.863858	CLUS3
CLUS43	F569	0.842291	CLUS97
CLUS43	F538	0.826999	CLUS97
CLUS43	F568	0.734813	CLUS20
CLUS43	F571	0.778924	CLUS3
CLUS43	F539	0.769454	CLUS3
CLUS43	F537	0.796731	CLUS97
CLUS43	F331	0.062634	CLUS92

Figure 15. Results from Single-Stage Clustering

Therefore the information for the cluster or variable structure will not be lost or alternated significantly by using the two-stage method. A small loss of information using the two-stage method instead of the single-stage method could be offset by the advantage of performance in run time. We have done a run-time comparison between two clustering methods with several sizes of data. We duplicated the real example data to get more variables, but we kept the same number of observations. We used a PC that has an Intel Core 2 Duo CPU at 1.99GHz and 3GB RAM for the benchmarking test. The result is summarized as follows.

Table 2, Comparison of Run Time between Single-Stage and Two-Stage Clustering

Number of Variables	Variable Clustering	Two-Stage Variable Clustering
617	2 Min. 27 Sec.	1 Min. 46 Sec.
1234(= 617 x 2)	8 Min. 14 Sec.	3 Min. 33 Sec.
1851(= 617 x 3)	25 Min. 41 Sec.	6 Min. 15 Sec.
2468(= 617 x 4)	1 Hr. 14 Min. 46 Sec.	10 Min. 37 Sec.
4937(= 617 x 8)	8 Hr. 27 Min. 35 Sec.	1 Hr. 07 Min. 58 Sec.

The test table shows the two-stage method is much faster than the single-stage method. The two-stage method is scalable up to a certain point, but the single-stage method is not scalable at all. The non-scalability of the two-stage method results from the first stage, in which we calculate the distance matrix, so it could be improved by using a scalable algorithm for the distance measure. Overall, the two-stage method can reduce the run time dramatically for large data sets.

VARIABLE CLUSTERING AND PREDICTIVE MODELING

Using the variable-clustering variable clustering node, we can make two types of predictive models: a latent variable predictive model with cluster components, and a predictive model with selected best variables. We can also combine them for predictive modeling. As a default, the variable clustering node produces cluster components that are the first principal components from each cluster. The cluster components scarify the orthogonal property, but increase the interpretability of input space.

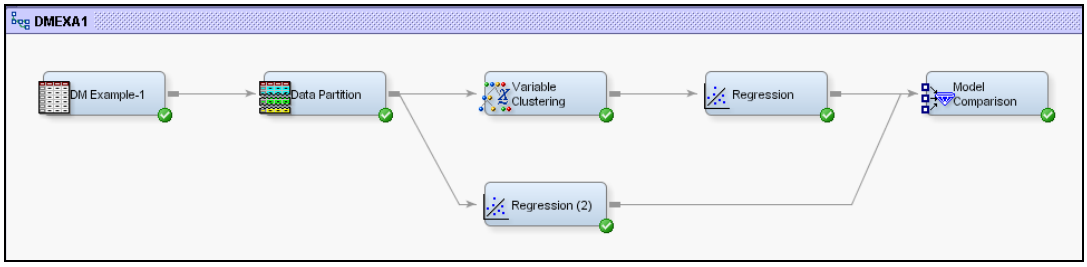


Figure 16. Predictive Modeling with Variable Clustering

The variable clustering node creates 11 cluster components from 43 interval variables, so the dimension reduction rate is one-fourth. The variable summaries before and after dimension reduction are shown below.

Variable Summary

Role	Measurement Level	Frequency Count
ID	NOMINAL	2
INPUT	INTERVAL	43
INPUT	NOMINAL	4
TARGET	BINARY	1

Variable Summary

Role	Measurement Level	Frequency Count
INPUT	INTERVAL	11
INPUT	NOMINAL	4
TARGET	BINARY	1

Figure 17. Variable Summaries

Because class variables are not included at the variable clustering, they are passed to the regression node without any changes. You can include class variables in the variable clustering process, but careful interpretation is required because dummy variables from one class variable could be clustered into different categories.

Name	Use	Report	Role	Level
PURCHASE	Yes	No	Target	Binary
Clus9	Default	No	Input	Interval
Clus10	Default	No	Input	Interval
Clus4	Default	No	Input	Interval
Clus7	Default	No	Input	Interval
Clus2	Default	No	Input	Interval
Clus8	Default	No	Input	Interval
Clus6	Default	No	Input	Interval
Clus3	Default	No	Input	Interval
Clus1	Default	No	Input	Interval
Clus11	Default	No	Input	Interval
Clus5	Default	No	Input	Interval
NTITLE	Default	No	Input	Nominal
GENDER	Default	No	Input	Nominal
APRTMNT	Default	No	Input	Nominal
TELIND	Default	No	Input	Nominal

Figure 18. Input Variables after Dimension Reduction

There is another way to use the variable clustering node for dimension reduction, which is a variable selection tool.

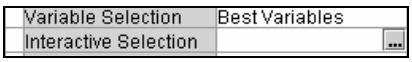
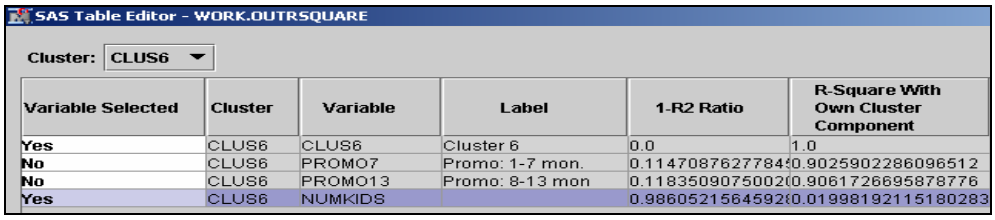


Figure 19. Property of Variable Selection

If you can choose the *Best Variables* option for the property of Variable Selection, the node will export the best variable per cluster based on the 1-R2 Ratio. The 1 -R2 Ratio gives the ratio of one minus the R2 with its own cluster component to one minus the R2 with the cluster component of next closest cluster. A small ratio indicates a good clustering. Notice that cluster components are always zero at the ratio because the numerator is zero. We can choose more variables or cluster components at a time by using the interactive selection editor. For example, the following cluster export cluster component (Variable = CLUS6) is the best representative variable for cluster 6. However, the NUMKIDS variable has a very low squared correlation with its cluster component, so you might add the variable to the predictive model with a slight chance of a collinearity problem. Alternatively, you can choose either PROMO7 or PROMO13 instead of cluster components because both promo variables are strongly related to the cluster component, and choosing actual variables increases the interpretability of predictive models.



Variable Selected	Cluster	Variable	Label	1-R2 Ratio	R-Square With Own Cluster Component
Yes	CLUS6	CLUS6	Cluster 6	0.0	1.0
No	CLUS6	PROMO7	Promo: 1-7 mon.	0.11470876277845	0.9025902286096512
No	CLUS6	PROMO13	Promo: 8-13 mon	0.11835090750026	0.9061726695878776
Yes	CLUS6	NUMKIDS		0.98605215645926	0.01998192115180283

Figure 20. Interactive Variable Selection Editor

For a comparison purpose, we used the regression node with and without the variable clustering node in the diagram. The following figure shows that the use of variable clustering increases model prediction accuracy and also provides dimension reduction, and better interpretability than regular principal components. The regression (2) model without variable clustering has a slight overfitting problem, which could result from multicollinearity or data redundancy, because the train misclassification is less than both validation and test ones. However, the model with variable clustering is fairly consistent in the misclassification rate over all data sets because even though the cluster components are not orthogonal, using cluster components treats a certain degree of multicollinearity.

Fit Statistics
Model selection based on _TMISC_

Selected Model	Model Node	Test: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error	Valid: Misclassification Rate
Y	Reg	0.40948	0.22750	0.40510	0.24152	0.41186
	Reg2	0.42132	0.21527	0.36178	0.25095	0.44407

Figure 21.Model Comparison

Because the use of two-stage variable clustering at predictive modeling is the same as that of the single-stage variable clustering, we omitted a predictive modeling example that uses two-stage variable clustering in this paper.

CONCLUSION

We showed two types of variable clustering and a combined method of two-stage variable clustering that uses variable clustering that is based on a distance matrix and a noble variable clustering that uses factor analysis with a rotation. Through examples and a benchmarking test, we showed the proposed two-stage method is appropriate to handle very large data sets with a minimum loss of variable structure information, and it is much faster than the single-stage method. Cluster components for predictive modeling are also obtained from each cluster. They are non-orthogonal principal components, but they increase their interpretation and remedy for multicollinearity in the predicted models. We observed a non-scalable issue at the two-stage method with more than 5,000 variables. Because the issue mainly comes from the calculation of all pair-wise distances at the first stage, a scalable algorithm for distance matrix calculation could improve the two-stage variable clustering. Further research on this issue is required.

REFERENCES

1. SAS Institute Inc. 1998. *SAS/STAT® User's Guide, Version 8*, Cary, NC: SAS Institute Inc.
2. Cerrito, Patricia B., and Vicki Viar. 1999. "An Investigation of Quality-of-Life Issues in Patient Treatment Using Enterprise Miner™." In *Conference Proceedings of the Twenty-Fourth Annual SAS® Users Group International Conference*. Cary, NC: SAS Institute Inc.
3. Nelson, Bryan D. 2001. "Variable Reduction for Modeling using PROC VARCLUS." In *Conference Proceedings of the Twenty-Sixth Annual SAS® Users Group International Conference*. Cary, NC: SAS Institute Inc.
4. D'Agostino, Ralph B., and Heidy K. Russell. 1998. "Cluster Analysis, Variables." In *Encyclopedia of Biostatistics*. vol. 1, 731-739.
5. Sanche, Robert, and Kevin Lonergan (2006). "Variable Reduction for Predictive Modeling with Clustering," *Casualty Actuarial Society Forum*, Winter 2006, 89-100.

ACKNOWLEDGMENTS

We thank Warren Searle for his helpful discussion of the PROC VARCLUS algorithm.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Taiyeong Lee
SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513
919-531-2186
Fax: 919-677-4444
E-mail:Taiyeong.Lee@sas.com

David Duling
SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513
919-531-5267
Fax: 919-677-4444
E-mail:David.Duling@sas.com

Song Liu
SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513
919-531-1355
Fax: 919-677-4444
E-mail:Song.Liu@sas.com

Dominique Latour
SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513
919-531-6312
919-677-4444
E-mail:Dominique.Latour@sas.com