Title: Analyzing messy and wide data on a desktop
Presenter: John Sall
Session: SAS2082

Analysis work can be challenged by data that is too wide, too full of miscodings, outliers, or holes, or by funny data types. Wide data, in particular, has many challenges, requiring the analysis to adapt with different methods. Making covariance matrices with billions of elements is just not practical.

Keywords: power SVD imputation, minimum-edit-distance recoding, outlier analysis, wide discriminant analysis.

Messy Data

On Aug 17, 2014, The Wall Street Journal featured an article titled, "For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights." Besides pitching the hot phrase *big data*, the article revealed a dirty little secret: data analysts spend much of their time preparing the data, at least the good analysts do.

So what can we do to make the cleanup faster, easier and more effective?

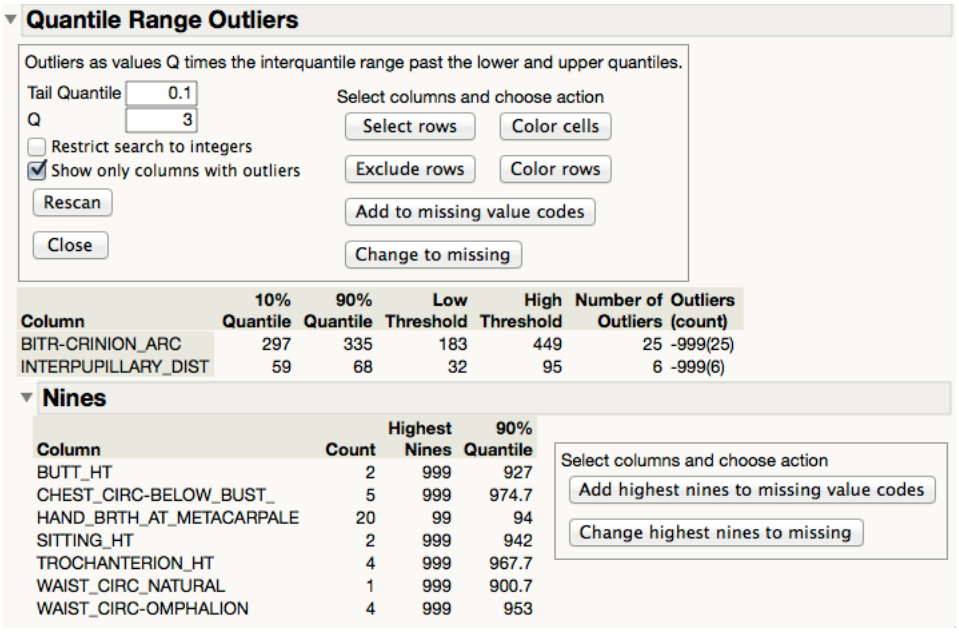
First, take categorical data. The big problem here is miscoding, which happens when people enter categories with a variety of spellings, punctuations and abbreviations. With a Recode facility, manual corrections can be made fairly easily, but what if you have hundreds or thousands of categories – some automation might save a lot of time. It turns out that a good criterion for comparing the closeness of two strings is ***edit distance*** (a variation of which is called *Levenshtein* distance). Edit distance is the minimum number of edit operations needed to turn one string into another. By looking at the edit distance and picking a criterion value, all the suggested combining of strings is done automatically; you just need to check it for groupings of different categories mistakenly combined. You will also still need to group categories that were the same but had a large edit distance.

For example, consider the customer field in the "Cylinder Bands" data that can be downloaded from the UCI Machine Learning laboratory [https://archive.ics.uci.edu/ml/datasets/Cylinder+Bands]. When you import this data, you will find that there are 83 unique values of Customer. However, the number of actual customers is much smaller than that since there are multiple codings of the same customer. For example, the customer Abbey Press is coded in three ways: "ABBEY", "ABBEYPRESS", and "ABBYPRESS". Edit distance captures two of these values, and you must then manually add the third one to the group. For Hanoverhouse, it grouped five of the six values.

Count	Old Values (84)	New Values (64)
9		
4	ABBEY	ABBEY
2	ABBYPRESS ▼	ABBYPRESS
2	ABBEYPRESS	
1	ADCO	ADCO
24	AMES	AMES
8	AUSTADS	AUSTADS
4	BELK ▼	BELK
2	BELKS	
2	best	best
9	BESTPROD ▼	BESTPROD
2	bestprod	
2	BRENDLS	BRENDLS
2	BURDINES	BURDINES
2	CAMPINGWORLD	CAMPINGWORLD
2	CASLIVING	CASLIVING
2	CASUALLIVING	CASUALLIVING
2	CENPURCH	CENPURCH
9	CHILDCRAFT	CHILDCRAFT
2	CHILDWORLD	CHILDWORLD
1	colorfulimage	colorfulimage
5	COLORTILE ▼	COLORTILE
2	COLORTIL	
2	COMPETITIVEDGE	COMPETITIVEDGE
2	CVS	CVS
7	DOWNS	DOWNS
4	DUNNS	DUNNS
9	ECKERD ▼	ECKERD
7	ECKERDS	
2	eckerd	
1	eckerds	eckerds
2	EXXON	EXXON
1	GALLS	GALLS
1	GLOBAL	GLOBAL
1	GLOBALEQUP	GLOBALEQUP
6	GUIDEPOSTS	GUIDEPOSTS
2	GURNEY	GURNEY
2	HANHOUSE	HANHOUSE
4	HANOVERHOUSE ▼	HANOVERHOUSE
2	HANOVERHSE	
2	HANOVRHOUS	
1	HANOVRHOUSE	
1	hanoverhouse	
8	HILLS	HILLS
4	homeshop ▼	homeshop
2	HOMESHOP	
2	HOMESHOPPING ▼	HOMESHOPPING
1	homeshopping	
3	JAMESWAY	JAMESWAY
4	JCP ▼	JCP
1	jcp	
2	JCPENNY	JCPENNY

For continuous data, the challenge is different – bad values are detectable when they are extremes relative to the range, i.e., outliers. The most straightforward approach is to get some tail quantiles and then measure some scale of the interquantile range past the tail quantiles.

The Ansur anthopometry data provides a good illustration, since it measures 131 body measurements. A scan revealed two columns with a value of -999 more than three interquantile ranges past the 10 percent and 90 percent tails. In this case, it is obvious that they are missing value codes, and it is easy to specify the code or change them into missing values. We also check for high-nines, since these are often used as missing value codes, but all the high-nines are not far from the upper tail of the data and are thus not likely to represent missing values.



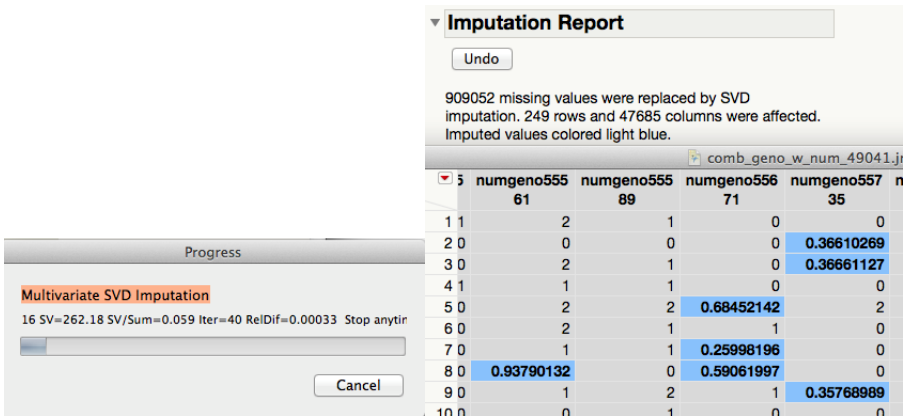
There are other outlier detection strategies implemented that use robust estimates, as well as two multivariate outlier detection methods: one that uses robust estimates and another that uses k-nearest-neighbors.

Missing Data

It is not uncommon to do multivariate analysis on many columns, only to receive an alert message saying that there is no data, since most multivariate techniques throw out any row that has a missing value in one of the columns used. It is useful to explore the data first with respect to the missingness of the data, and then have options to impute the values. A very simple multivariate normal imputation can be done with regression calculations using the non-missing variables to predict the missing ones. This, however, becomes problematic when you have many thousands of columns.

One approach to make imputation practical is to use a special transform on the data, the singular value decomposition. This approach approximates the data matrix by a multiplication of three matrices with special properties that goes through a smaller number of dimensions.

For example, below is a data table with 49,041 columns of gene expressions across 249 rows. After about 30 seconds, the variance extracted in the 16th dimension is only about 6 percent of the sum, so we stop it and then multiply the rows and columns of the singular vectors and values to impute for a missing value. This method would not be practical using normal (least squares) techniques, because the pairwise covariance matrix would have 1.2 billion unique elements.



Wide Problems

The previous problem with 49,000 columns illustrates an important point – very wide problems need a different computational strategy than one used for narrow problems. Generally, we will use an SVD rather than composing a covariance matrix.

We have done this for two multivariate platforms, Principal Components and Discriminant Analysis. All the ingredients are there in the SVD without calculating the covariance matrix.

For example, consider the breast cancer gene expression data from the Microarray Quality Control consortium, with 10,787 gene expressions on 230 samples across several hormone status levels. In less than six seconds, we can do a full discriminant analysis, one that is closely competitive on cross-validated misclassification rate with the best methods found for this benchmark problem.

Score Summaries

Source	Count	Number Misclassified	Percent Misclassified	Entropy RSquare	-2LogLikelihood
Training	179	6	3.35196	0.44007	92.3894
Validation	51	3	5.88235	0.75509	

Training

Actual HER2_Status	Predicted N	Predicted P
N	146	2
P	4	27

Validation

Actual HER2_Status	Predicted N	Predicted P
N	40	2
P	1	8

We can even save much more efficient formulas. The problem with tens of thousands of columns is that the scoring formulas repeatedly involve tens of thousands of columns. We form a single data column with all 49,000 values in one cell, and then use matrix arithmetic to do all the scoring formulas, resulting in much faster scoring, as well as eliminating most of the overhead due to formula ordering logic.

Discrim Data Matrix	Discrim Prin Comp	SqDist[0]	SqDist[N]	SqDist[P]	Prob[N]	Prob[P]	Pred HER2_Status
[7.0390625, 5.1083…	[1.216586445711…	176.55934385	175.80405405	189.49775635	0.99894	0.00106	N
[6.421875, 5.01562…	[1.200253551373…	175.1889649	175.80405405	181.58492306	0.94737	0.05263	N
[7.181640625, 5.08…	[1.058564052060…	175.02461935	175.80405406	180.63596001	0.91804	0.08196	N
[9.6328125, 5.0830…	[1.109189148316…	140.9827232	136.58929291	171.29032258	1.00000	0.00000	N
[8.552734375, 5.34…	[1.275206715209…	173.59148699	175.80405405	172.36077641	0.15166	0.84834	P
[7.4462890625, 4.9…	[1.011471680754…	175.73715711	175.80405405	184.75029095	0.98872	0.01128	N
[5.634765625, 5.10…	[1.278024226985…	175.0527442	175.80405405	180.79835839	0.92394	0.07606	N

This analysis uses the recently released JMP® 12 from SAS.