Session 1

Exercise Answers

Exercise B.

B. The general probability model for categorical variables

1. For the 3-class model, how many probability parameters are there? How many unconditional probabilities (associated with the size of each latent class)? How many conditional probabilities?

   Answer: The Profile output contains the probability parameters. There are 3 unconditional and 30 conditional parameters shown.

2. Regarding the K unconditional probabilities, since they sum to 1, only K-1 are distinct - the last one can always be computed from the others. In total, for the 3-class model, how many distinct parameters are there? Does this agree with the number reported under the Npar column as shown in Figure 7-9 of LG Tutorial 1 (see page 41)?

   Answer: Two of the indicators contain 2 categories each and the other two indicators contain 3 categories each, for a total of 10 conditional probabilities, for each of the 3 classes. Of these 10, 4 (1 for each of the 4 indicators) are redundant since they can be obtained from the conditional probabilities associated with the other categories. Thus, the total number of 'distinct' conditional probability parameters is 3(10-4) = 18, and the total number of distinct parameters is 20 (18 + 2 distinct unconditional probabilities).

Exercise C.

C. Determining the number of classes/clusters

1. What criteria do you use to determine the number of classes?

   Answer: There is no single right answer to this question. Since the various statistical criteria do not always agree with each other on the number of classes, there is room for a subjective element in determining the number of classes. Thus, it makes sense to select a model that yields classes that are interpretable and meaningful. For example, you may choose a 4-class model over a 5-class model if the 5th class is very small – say it contained less than 1% of the cases. On the other hand, rather than accepting the 4-class model (which probably combines this 5th

class with one of the other classes) the 5th class might identify an outlier which you may want to exclude rather than combining it with another class.

With many indicators, one may find that the BIC does not start decreasing until many more latent classes are specified than are meaningful from a practical/substantive perspective -- perhaps several hundreds of segments would need to be specified before the BIC would start to decline. This simply means that from a pure statistical perspective, adding an additional class beyond say 50 to 51 (or beyond 123 to 124, etc.) provides some additional discriminatory power that is justified by the data. However, it is a good idea to take into account practical criteria along with the statistical criteria.

If you are satisfied with the number of classes based on some criteria, requesting this number of classes in Latent GOLD or other latent class routine which employs maximum likelihood estimation will almost always provide a 'better' solution than that given by K-Means, hierarchical or other clustering algorithm with the same number of segments because you are getting the most likely (maximum likelihood) solution under the specified model. By 'better', I mean one that would be judged to be more meaningful from a substantive perspective.

Exercise D.

D. Fit measures, model specification and selection strategies

1. Reproduce the table shown in Table 5 of the SAGE article (page 21) for the 1-class model H0 by computing the Pearson chi-square test for independence using the raw data. Hint: Be sure to divide by the appropriate number of degrees of freedom. Do some of the 4 variables appear to be independent? Which ones?

   Answer: For the 1-class model, the BVRs associated with {AC} and {BC} are both very small suggesting that (C) Understanding is independent of both (A) Purpose and (B) Accuracy.

   Note: The BVRs reported using the default values of Latent GOLD will be slightly lower than that obtained by dividing the Pearson chi-squared by the number of degrees of freedom (DF) because of the default value for the Bayes constant (which may be set in the Technical Tab of Latent GOLD) is '1' rather than '0'. The Bayes constant will be discussed later. For example, the BVR associated with the variable pair (PURPOSE, ACCURACY) is 61.64, which is close to the Pearson chi-squared divided by DF (123.385 / 2 = 61.69). If the Bayes constant for categorical variables is changed from the '1' to '0' prior to estimating the model, the BVR becomes 61.69.

Exercise E.

E. Classifying cases into latent class segments

1. If you have access to SPSS, use the ClassPred Tab to obtain the standard classification output to the file 'data3.sav' as indicated in Figure 7-20 of LG Tutorial 1 (page 50). If you do not have access to SPSS, use the Edit Copy command to copy the standard classification output (shown in Figure 7-19 on page 49) to the Clipboard and paste it into Excel. Then sort by 'Modal'.

2. Compute the frequency distribution for the modal assignment class and confirm that it is the same as shown in Figure 7-10 (LG Tutorial 1, page 42) - 805(1), 178(2), and 219(3). Why does this distribution not match the cluster sizes as reported in the Profile Output in Figure 7-14 of LG Tutorial 1(page 45).

   Answer: The discrepancy is due to misclassification error caused by modal assignment. The relationship between modal assignment and probabilistic assignment is shown in the classification table in Figure 7-10 in the Tutorial (page 42).

Exercise F.

F. Interpreting Latent GOLD output

1. The Tri-plot makes clear that the 3-class model may be considered to be a 2-dimensional model. How many dimensions are associated with a 4-class model? Is the tri-plot meaningful in the case of more than 3 classes?

   Answer: There are 3 dimensions associated with a 4-class model. The number of dimensions is determined by the number of distinct contrasts that is possible between the classes. For a 2-class model, only 1 contrast is possible (class 2 vs. class 1). Thus, a 2-class model is uni-dimensional. In the case of a 3-class model, selecting class 1 as the reference, we have 2 distinct contrasts -- 2 vs. 1 and 3 vs. 1, thus the 3-class model is a 2-dimensional model. In the case of a K-class model, there a K-1 dimensions.

   The triplot displays the results in only 2 dimensions. The tri-plot in Latent GOLD associates 2 selected classes with the lower vertices, and utilizes all other classes as the upper vertex of the triangle. In the case of a model with 3 or more dimensions, that triplot represents a simplification resulting in loss of information by aggregating 2 or more classes, displaying them together in a single vertex.

Exercise G.

G. Example from survey analysis

1. From a substantive perspective, how might you interpret the results as displayed in the Tri-plot (see LG Tutorial 1, Figure 7-17, page 48)? In particular, the categories of UNDERSTANDING appear to trace out a horizontal dimension in the tri-plot, while the categories of the other variables seem to trace out more of the vertical dimension.

   Answer: The categories of PURPOSE and ACCURACY trace the vertical dimension while the categories of COOPERATE are associated with both dimensions. Thus, one interpretation is that those judged to be impatient/hostile may be so because of failure to understand the survey questions (the horizontal dimension) or because they do not believe in the value of surveys as indicated by their responses to PURPOSE and ACCURACY (the vertical dimension).

   Note: The fact that the categories of UNDERSTANDING trace out a horizontal dimension in the tri-plot while the categories of PURPOSE and ACCURACY trace out a vertical dimension suggest that UNDERSTANDING is independent of both PURPOSE and ACCURACY. The independence between these 2 variable pairs was confirmed in exercise D1. Generally speaking, such orthogonality between 2 variables in the tri-plot implies independence.

Exercise H.

H. Including covariates in LC models

1. When do you think covariates should be treated as active and when inactive?

Answer: When many covariates are available, it is generally not advisable to make all these covariates active, because the main-effects-only assumption may not hold true. In such a case, by imposing this restriction, it is possible that the class sizes may change substantially from what they turn out to be when no active covariates are specified. Since this change is due to the forcing the main-effects-only assumption to hold, the class size estimates may be very misleading. When covariates are small in number, it is largely a matter of personal preference whether or not to make them active. Note that by default an additional assumption made is that active covariates are conditionally independent of the indicators given the classes.

Exercise I.

I. Boundary, identification and local solution issues; Bayes constants

1. Re-estimate the 3-class model estimated earlier in tutorial #1. Now, change the technical parameter setting for the 'Bayes Constants for Categorical Variables' from '1' to '0' and estimate a new model. (You will find the 'Bayes Constants' settings labeled in the upper right portion of the Technical Tab.)

Notice that the L-squared statistic is now slightly better than the original model (L^2 = 21.8920 vs. the original value of 22.0872). An Estimation Warning message is produced along with an Iteration Output file. At the bottom of the Iteration Output is a message saying that 2 boundary solutions were encountered. Go to the Profile output. Where are the 2 boundary solutions that were encountered? How do these estimates compare to the corresponding estimates obtained in the original model?

Answer: The 2 boundary solutions (extreme estimates of '1' and '0' for certain conditional probabilities) are highlighted in yellow in the table below.

	Cluster1	Cluster2	Cluster3
Cluster Size	0.6183	0.2087	0.173
Indicators			
purpose			
good	0.8884	0.9119	0.1439
depends	0.0530	0.0716	0.2245
waste	0.0586	0.0165	0.6316
accuracy			
mostly true	0.6131	0.6481	0.0321
not true	0.3869	0.3519	0.9679
understa			
good	1	0.3190	0.7532
fair/poor	0	0.6810	0.2468
cooperat			
interested	0.9438	0.6901	0.6413
cooperative	0.0562	0.2554	0.2560
impatient/hostile	0	0.0545	0.1027

How do these estimates compare to the corresponding estimates obtained in the original model?

Answer: In the model with Bayes = 1, the corresponding conditional probability parameter estimates are less extreme, avoiding probability estimates of '1' and '0'.

	Cluster1	Cluster2	Cluster3
Cluster Size	0.6169	0.2038	0.1793
Indicators			
purpose			
good	0.8905	0.9157	0.1592
depends	0.0524	0.0706	0.2220
waste	0.0570	0.0137	0.6189
accuracy			
mostly true	0.6148	0.6527	0.0426
not true	0.3852	0.3473	0.9574
understa			
good	0.9957	0.3241	0.7532
fair/poor	0.0043	0.6759	0.2468
cooperat			
interested	0.9452	0.6879	0.6432
cooperative	0.0547	0.2583	0.2559
impatient/hostile	0.0001	0.0538	0.1009

Next, change the Bayes constant from '0' to '2' and estimate the model. Compare the profile output in these two models.

Answer: Bayes = 2 smooths the parameter estimates further. As shown in the following table, the highlighted estimates are now further away from '0'.

	Cluster1	Cluster2	Cluster3
Cluster Size	0.6144	0.2020	0.1835
Indicators			
purpose			
good	0.8918	0.9169	0.1714
depends	0.0520	0.0703	0.2198
waste	0.0562	0.0128	0.6088
accuracy			
mostly true	0.6161	0.6566	0.0483
not true	0.3839	0.3434	0.9517
understa			
good	0.9921	0.3343	0.7534
fair/poor	0.0079	0.6657	0.2466
cooperat			
interested	0.9468	0.6857	0.6450
cooperative	0.0530	0.2612	0.2555
impatient/hostile	0.0002	0.0531	0.0995

Notice that as the Bayes constant is increased, the extreme parameter estimates become less extreme. The greater the value of the Bayes constant, the greater the weight that is placed on a conservative null model ('prior distribution') which specifies that all variables are mutually independent. Use of the default Bayes constant of 1 provides a fairly small weight for this 'prior' distribution.

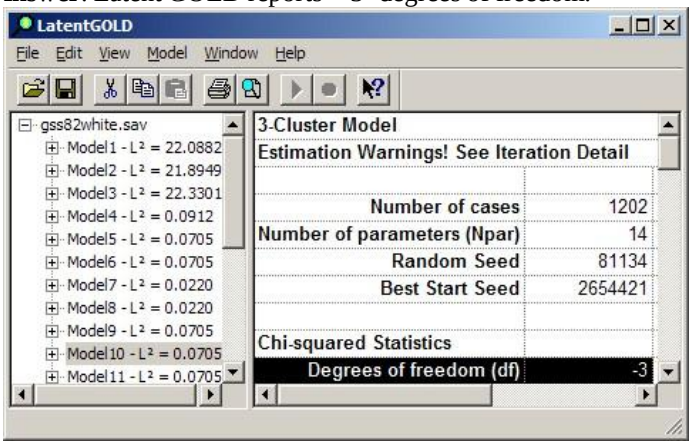
2. Return to the 3-class model estimated in Exercise B1 when the Bayes constant was set to 0. Open the Variables Tab, remove the variable COOPERATE from the model and estimate it. Again, you will get an Estimation Warning Message. Estimate the model once again. Do you get the same L-squared value? If not, re-estimate it again until you get the same L-squared value at least twice. Examine the 'Parameters' Output for models having the same L-square value. Notice that some of the parameter estimates are different! This is an indication that these parameter estimates are not identified.

For this exercise, you may obtain an L-squared value of .0705 (which is actually an unidentified 'local' solution), or .0411 which is the unidentified 'global solution' (At least I believe that it is the global solution. When local solutions exist, it is never 100% certain that the solution you obtain is a global solution.) You may also encounter other L-squared values associated with other local solutions when estimating this model. See the section on 'Local solutions' above.)

Now, repeat the exercise after restoring the Bayes Constant to its default value of '1'. Notice that for models having the same L-squared value, the parameters estimates no longer are different. That is because the information provided by non-zero Bayes constant is sufficient to uniquely identify the model.

How many degrees of freedom are associated with this model?

Answer: Latent GOLD reports '–3' degrees of freedom.



3. If you estimate and re-estimate a model several times and always get the same L-squared value and always the same parameter estimates, and the output is very interpretable from a substantive perspective, can you be comfortable with your interpretation?

Answer: Generally, 'Yes'.

What if you notice that the degrees of freedom are negative?

Answer: This indicates that 1 or more parameter estimates are not identified. The fact that you get the same parameter estimates may be due to the non-zero Bayes constant used. You can experiment with different values for the Bayes constant to see how the results change. If df = -1, it may be that only 1 parameter is not identifiable. In such a case, comparing estimates obtained from repeated estimations, you will notice that only estimates for a single parameter change, the others remaining the same. In such a case, it is permissible to make inferences regarding the parameters that are identified.

Exercise J/L.

J/L. Exercise: Example with Diabetes data

1. Read the diabetes example in CAMBRIDGE, pages 14-18 and SAGE section 4.3.

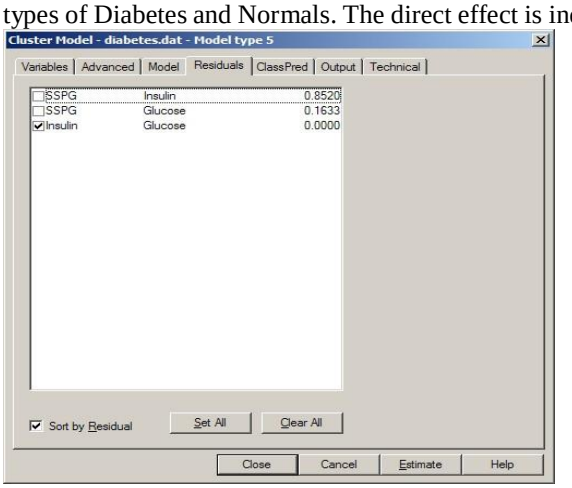
Download the associated data file diabetes.dat and diabetes.lgf file for these data . Load the .lgf file and re-estimate all the models in the .lgf file (Note: these are 3-class models only).

After estimating a model, double click on that model and click the Residuals Tab. Here you will see the bivariate residuals associated with each pair of indicators, sorted from high to low. A checkmark preceding an indicator pair indicates that a direct effect parameter for that pair has been included in the model. In the Model tab, you will see which (if any) of the effects are specified as class independent. Which model do you think is best? What is your criteria?

Answer: According to the BIC, Model Type 5 is preferred among these 3-class models.

		LL	BIC(LL)	Npar	Class.Err.
1. Class-dependent and full VarCov matrix	3-Cluster	-2308.6431	4761.6116	29	0.0547
2. Class-independent and full VarCov matrix	3-Cluster	-2419.1089	4922.8223	17	0.0025
3. Class-dependent and diagonal VarCoV matrix	3-Cluster	-2366.9233	4833.3813	20	0.0376
4. Class-independent and diagonal VarCoV matrix	3-Cluster	-2464.783	4999.2402	14	0.0038
5. Class-dependent VarCoV matrix with only Glucose-Insulin covariance	3-Cluster	-2320.5748	4755.6144	23	0.0609
6. Class-independent VarCoV matrix with only Glucose-Insulin covariance	3-Cluster	-2440.8545	4956.36	15	0.012

The parameters for this model are those shown in Table 2 of the Cambridge chapter. This model relaxes the assumption of local independence by including a direct effect between Insulin and Glucose. This suggests that the correlation between Insulin and Glucose is not relevant to the discrimination between the 3 classes – the 2 different types of Diabetes and Normals. The direct effect is indicated by a check-mark in the Residuals Tab.



Add the true diagnosis – the variable TRUE -- as an inactive covariate in each of these models.
Examine the Profile and ProbMeans output to see which model most closely relates the latent classes to the desired true states.

Answer: Class 1 refers to Chemical Diabetes, Class 2 to Normals, and Class 3 to Overt Diabetes. For Model Type 5, the Profile output shows that among class 1, 88% were diagnosed as having Chemical Diabetes, among class 2, 66% were diagnosed as being normal, and among class 3, 98% were diagnosed as having Overt Diabetes.

	Cluster1	Cluster2	Cluster3
Cluster Size	0.5391	0.2696	0.1913
Indicators			
Glucose			
Mean	91.2	104.0	234.8
Insulin			
Mean	359.2	495.1	1121.1
SSPG			
Mean	163.1	309.4	77.0
Covariates			
TRUE			
1	0.12	0.66	0.02
2	0.88	0.19	0.00
3	0.00	0.15	0.98

The Covariate portion of the ProbMeans output for Model Type 5 shows that among all Normals (TRUE = 1), 72% were classified into class 2 while 27% were 'mistakenly' classified into class 1 (the 'Chemical Diabetes' class). Hence, the false positive rate is 28%. The false negative rate is lower.

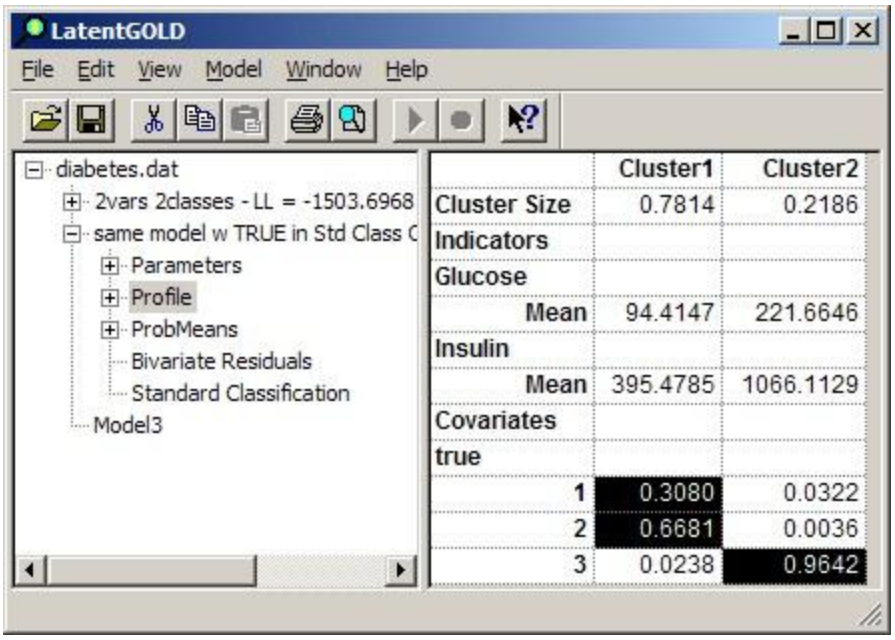
	Cluster1	Cluster2	Cluster3
Overall	0.5391	0.2696	0.1913
Indicators			
Covariates			
TRUE			
1	0.27	0.72	0.01
2	0.90	0.10	0.00
3	0.00	0.18	0.82

2. Re-estimate model type 5, requesting the posterior membership probabilities (Classification - Posterior) be output to a file. Then open the newly created outfile and use the new Step 3 option to obtain the scoring formula that can be used to score new cases as a function of the 3 indicators. Hint: Since model type 5 does not assume the variances and covariances to be equal within each of the 3 latent classes, a quadratic function must be specified in order to obtain an R^2=1 (i.e., to perfectly reproduce the posterior membership probabilities). For assistance, see: 'Step 3 Tutorial 2'. Which quadratic terms entered into the model have non-zero coefficients? What is the formula for the posterior membership probabilities as a function of the 3 indicators?

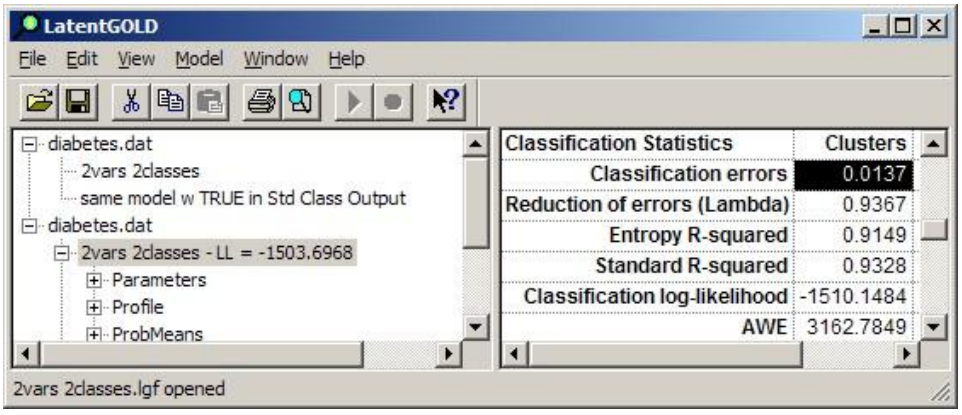
Answer: The squared terms for each of the 3 indicators (Y1, Y2, Y3) = (SSPG, INSULIN, GLUCOSE) and the 'Insulin * Glucose' interaction term.
Prob[k | Y1, Y2, Y3) = exp[score(k | Y1, Y2, Y3)] / (exp[score(1 | Y1, Y2, Y3)] + exp[score(2 | Y1, Y2, Y3)] + exp[score(3 | Y1, Y2, Y3)]) for k = 1, 2, 3

3. Using only the 2 variables GLUCOSE and INSULIN, how well are you able to distinguish persons with overt diabetes from others using a 2-class model? How can you use Latent GOLD output to see that only 3 cases are misclassified?

Answer: If you include these 2 variables as indicators and estimate a 2-class Cluster model with a direct effect to account for within-class correlation, the Profile output shows that class 2 represents 21.86% of the population, and 96.42% of this class has 'overt diabetes' (TRUE = 3). On the other hand, only 2.38% of the remaining cases in the population (class 1) is estimated to have this type of diabetes. Thus, this model does very well in identifying persons with overt diabetes.

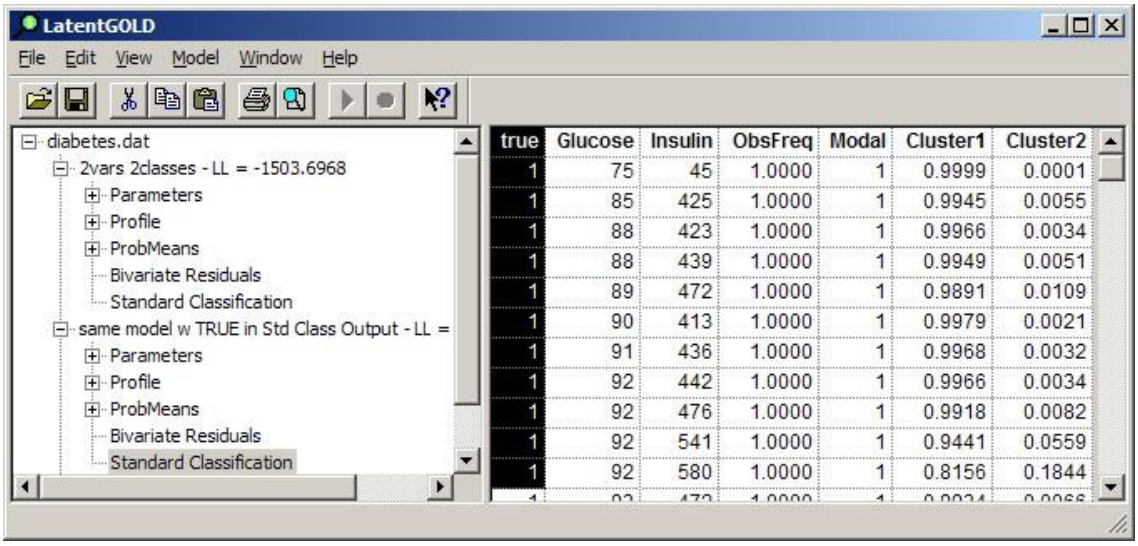


If modal assignment is used to classify cases into either class 1 or class 2 based on this model, the percentage of cases that would be expected to be misclassified would be only 1.37% as shown below.



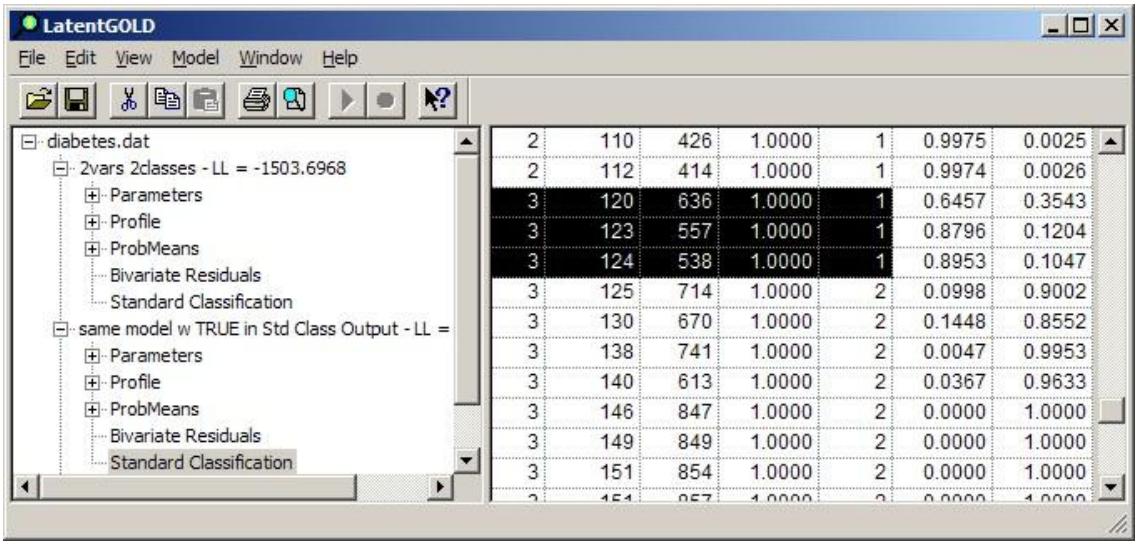
In a sample of 145 cases, we would therefore expect only 2 cases to be misclassified (145 x .0137 = 1.99). We find that in the current sample, there are in fact 3 cases that are misclassified as shown below.

To have the true diagnosis show up in the output but not influence the model parameters, you could specify the variable TRUE as an inactive covariate, or as an active covariate whose effect is restricted to zero. Both of these model specifications result in exactly the same model. The advantage of specifying TRUE to be an *active* covariate is that the true diagnosis shows up not only in the Profile output, but also the Standard Classification output.



true	Glucose	Insulin	ObsFreq	Modal	Cluster1	Cluster2
1	75	45	1.0000	1	0.9999	0.0001
1	85	425	1.0000	1	0.9945	0.0055
1	88	423	1.0000	1	0.9966	0.0034
1	88	439	1.0000	1	0.9949	0.0051
1	89	472	1.0000	1	0.9891	0.0109
1	90	413	1.0000	1	0.9979	0.0021
1	91	436	1.0000	1	0.9968	0.0032
1	92	442	1.0000	1	0.9966	0.0034
1	92	476	1.0000	1	0.9918	0.0082
1	92	541	1.0000	1	0.9441	0.0559
1	92	580	1.0000	1	0.8156	0.1844

This allows you to see that only 3 of the sample cases are misclassified (TRUE = 3, but MODAL = 1 instead of 2). The 3 misclassified cases are those who have been confirmed to have overt diabetes (TRUE = 3), but instead of being assigned to class 2, they were assigned to class 1.



2	110	426	1.0000	1	0.9975	0.0025
2	112	414	1.0000	1	0.9974	0.0026
3	120	636	1.0000	1	0.6457	0.3543
3	123	557	1.0000	1	0.8796	0.1204
3	124	538	1.0000	1	0.8953	0.1047
3	125	714	1.0000	2	0.0998	0.9002
3	130	670	1.0000	2	0.1448	0.8552
3	138	741	1.0000	2	0.0047	0.9953
3	140	613	1.0000	2	0.0367	0.9633
3	146	847	1.0000	2	0.0000	1.0000
3	149	849	1.0000	2	0.0000	1.0000
3	151	854	1.0000	2	0.0000	1.0000

4. (Optional): If you have access to SPSS, use the K-Means procedure (using the Analyze/Classify menu), specifying 2 clusters and requesting that cluster membership probabilities be saved. Confirm that 7 cases are misclassified. Now repeat the analysis after standardizing the variables to Z scores (using the Analyze/Description Statistics/Descriptives menu), check "Save standardized values". Are there more or less misclassifications?

Show that the latent class model is unchanged when Z scores are used.

Answer: The use of the Z-Scores in this example causes K-Means to actually do worse than when the original scale for the variables (Glucose and Insulin) are used. This is surprising because of the general recommendation to transform all variables to Z-Scores prior to using K-Means. In some cases, use of Z-Scores produces better results, while in other cases such as here, it makes it worse.

The general recommendation to use Z-Scores is due to the K-Means' implicit assumption that the within-class variances of all variables are equal. Since the true classes are unknown, the within-class variances are unknown, so it is not possible to make them equal. Short of that, Z-Scores are used to set the variances of all variables equal to each other overall. Sometimes, this standardization succeeds in making the within-class variances closer to being equal, while other times it actually makes within-class variances more different. In this example, the variances are substantially different with and without the use of Z-scores.

In this example, 3 assumptions implicit in the K-Means approach are violated. First, the within-class variances of the 2 variables differ. Secondly, the within-class correlation between the variables is non-zero. Third, the variance-covariance matrices are not the same between the 2 classes. Within the Overt Diabetes class the correlation is about .95; in the other it is about .6. Also, within the Overt class, the variances of both variables are much higher than within the other class -- regardless of whether the variables are standardized or not.

The fact that the overall correlation is quite high, as you point out, allows us to simplify the example to a single variable. This allows us to see more clearly why a correctly specified LC model does much better than K-Means here. Note that high values on Glucose (or Insulin) are more likely to be associated with the true Overt Diabetes class, and the standard deviation for Glucose (and Insulin) is much higher in the Overt class. With only a single variable, the only relevant issue for classification becomes what is the most appropriate cut-off point. K-Means uses distance as the single criterion. By ignoring the fact that the variance is higher in the Overt class, the cutoff chosen by the K-Means approach would be biased in the direction of being too high, which causes more true Overt cases to be misclassified.

This same misclassification bias can be shown to occur with misspecified LC models. If you use the model tab in Latent GOLD to restrict the error variances and/or covariances to be identical in the 2 classes ('cluster independent'), you will obtain misspecified models which cause many more of the Overt cases to be misclassified. However, if you use the BIC criteria to compare models, you will select as best the correctly specified model that allows the covariance matrices to differ across classes.

By allowing for such differences in the variances, the correctly specified LC model utilizes a probability-based distance measure (i.e., standard deviation units) to determine the best cut-off point. The K-Means approach utilizes raw distance without adjusting appropriately for within-class differences in the standard deviations. You can extend the optional exercise to use only a single variable. The appropriately specified LC model does quite well even with only a single variable.