randomLCA Examples

Ken Beath

September 1, 2014

1 Introduction

Following are two examples of using randomLCA for latent class analysis. Some aspects will certainly change but most code should still work. Two things that will change are the use of accessor functions and better labelling of results.

2 Latent Class

2.1 Model

The basis of latent class analysis is that each subject belongs to one of a finite number of classes, with each class described by a set of parameters that define the distribution of outcomes or manifest variables for a subject, a form of finite mixture model. For binary outcomes, the model for each class is

$$P\left(y_{i1}, y_{i2}, ..., y_{ik} | \, c\right) \;=\; \prod_{j=1}^{k} \pi_{cj}^{y_{ij}} \left(1 - \pi_{cj}\right)^{1-y_{ij}}$$

where

$$\begin{aligned} y_{ij} &= j\text{th binary outcome for subject } i \\ \pi_{cj} &= \text{probability of } j\text{th outcome equal to 1 for subject in class } c \\ k &= \text{number of outcomes.} \end{aligned}$$

An additional parameter that is required to be estimated is η_c, the probability of a subjects in class c.

A requirement for the estimates of the probabilities π_{cj} is that they be restricted to the interval zero to one, and η_c sum to one. This can be obtained using the following relations $\pi_{cj} = \frac{e^{a_{cj}}}{1+e^{a_{cj}}}$ and $\eta_c = \frac{e^{\theta_c}}{\sum_{\ell=1}^{C} e^{\theta_\ell}}$. Hence we estimate the π_{cj} and θ_c.

2.2 Example 1

This example demonstrates the fitting of data from Rindskopf and Rindskopf (1986), where latent class analysis is used to determine diagnostic classifications based on medical tests. Although this example is for medical data, the model is simply standard latent class so the methods can be applied to data from other areas.

A series of latent class models for 1 to 4 classes can be fitted using the commands

```
> myocardial.lca1 <- randomLCA(myocardial[,1:4],
+   freq=myocardial$freq,nclass=1)
> myocardial.lca2 <- randomLCA(myocardial[,1:4],
+         freq=myocardial$freq,nclass=2)
> myocardial.lca3 <- randomLCA(myocardial[,1:4],
+         freq=myocardial$freq,nclass=3)
```

The BIC values may be extracted from the fitted objects and are shown in Table 1.

```
> bic.data <- data.frame(classes=1:3,bic=c(BIC(myocardial.lca1),
+         BIC(myocardial.lca2),BIC(myocardial.lca3)))
```

classes	bic
1	524.7
2	402.3
3	421.1

Table 1: BIC by class.

Using BIC as a selection method, this selects the 2 class model, indicating a nice beakdown into diseased and nondiseased, which it is assumed represent those with and without myocardial infarction. The true nature of classes is always debateable.

Summary may be used to display the fitted results

```
> summary(myocardial.lca2)
```

```
  Classes      AIC      BIC    logLik penlogLik
        2 379.3958 402.2855 -180.6979 -180.7002
Class probabilities
Class  1 Class  2
  0.4578   0.5422
Outcome probabilities
         Q.wave History    LDH    CPK
Class  1 0.7668  0.7914 0.8279 1.0000
Class  2 0.0000  0.1951 0.0269 0.1955
```

Individual results may be obtained from summary, for example the outcome probabilities shown in Table 2.

```
> outcomep.data <- summary(myocardial.lca2)$outcomep
```

	Q.wave	History	LDH	CPK
Class 1	0.767	0.791	0.828	1.000
Class 2	0.000	0.195	0.027	0.196

Table 2: Outcome Probabilities.

This gives some interesting information. In Class 2, those without myocardial infarction, will have abscence of Q.wave but in those with myocardial infarction it will only be present in 76.7%. The class probabilities can be obtained as `myocardial.lca2$classp` of 0.46 and 0.54 for Class 1 and 2 respectively.

One aspect of latent class is that no subject is uniquely allocated to a given class, although in some cases a subject may have an extremely high probability.

The posterior class probs can be obtained as

```
> classprobs <- post.class.probs(myocardial.lca2)
```

with results shown in Table 3. This shows subjects with 3 or 4 positive tests to be strongly classified as having myocardial infarction, and even some with 2, depending on which to to be well classified. Having only one positive test makes it unlikely that it is myocardial infarction.

Q.wave	History	LDH	CPK	Freq	Class 1	Class 2
1	1	1	1	24	1.000	0.000
0	1	1	1	5	0.992	0.008
1	0	1	1	4	1.000	0.000
0	0	1	1	3	0.889	0.111
1	1	0	1	3	1.000	0.000
0	1	0	1	5	0.419	0.581
1	0	0	1	2	1.000	0.000
0	0	0	1	7	0.044	0.956
0	0	1	0	1	0.000	1.000
0	1	0	0	7	0.000	1.000
0	0	0	0	33	0.000	1.000

Table 3: Class Probabilities.

Outcome probabilities are shown in Figure 1.

```
> trellis.par.set(col.whitebg())
> print(plot(myocardial.lca2,type="l",xlab="Test",
+   ylab="Outcome Probability",scales=list(x=list(at=1:4,
+   labels=names(myocardial)[1:4]))))
```

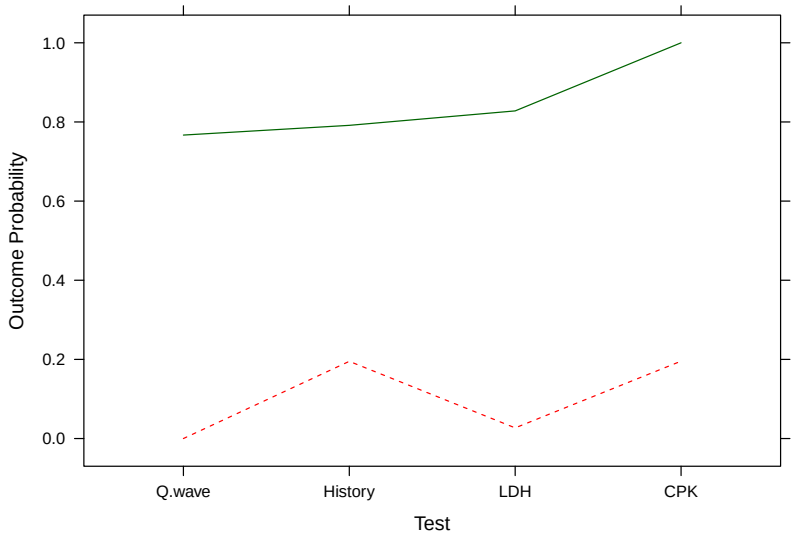

Figure 1: Outcome probabilities for 2 Class Latent Class model.

2.3 Example 2

This example shows the fitting of the dentistry data from Qu et al. (1996). The data consists of the results of five dentists evaluating x-rays for presence or absence of caries. As there is no gold standard, the latent class method is to assume two classes, diseased and non-diseased which are identified from the data.

A series of latent class models for 1 to 4 classes can be fitted using the commands

```
> data(dentistry)
> dentistry.lca1 <- randomLCA(dentistry[,1:5],
+         freq=dentistry$freq,nclass=1)
> dentistry.lca2 <- randomLCA(dentistry[,1:5],
+         freq=dentistry$freq,nclass=2)
> dentistry.lca3 <- randomLCA(dentistry[,1:5],
+         freq=dentistry$freq,nclass=3,quadpoints=31)
> dentistry.lca4 <- randomLCA(dentistry[,1:5],
+         freq=dentistry$freq,nclass=4,quadpoints=41)
```

The BIC values may be extracted from the fitted objects and are shown in Table 4. This indicates the presence of 3 classes. A possible interpretation is that there is a class of subjects with moderate disease, or the alternative of heterogeneous disease which will be covered in the next section. Outcome probabilities are shown in Figure 2 and for the 2 class model in Figure 3.

```
> bic.data <- data.frame(classes=1:4,bic=c(BIC(dentistry.lca1),
+         BIC(dentistry.lca2),BIC(dentistry.lca3),BIC(dentistry.lca4)))
```

classes	bic
1	17531.1
2	15021.6
3	14962.9
4	15000.0

Table 4: BIC by class.

The 2 Class results can be interpreted as a diagnostic test. Important results for diagnostic testing are the sensitivity and specificity for each test. The sensitivity is the probability of the test correctly identifying the subject as diseased given that the subject is diseased. In classical diagnostic testing the "true" status of a subject is known through use of a "gold standard"

```
> trellis.par.set(col.whitebg())
> print(plot(dentistry.lca3,type="l",xlab="Dentist",
+     ylab="Outcome Probability"))
```

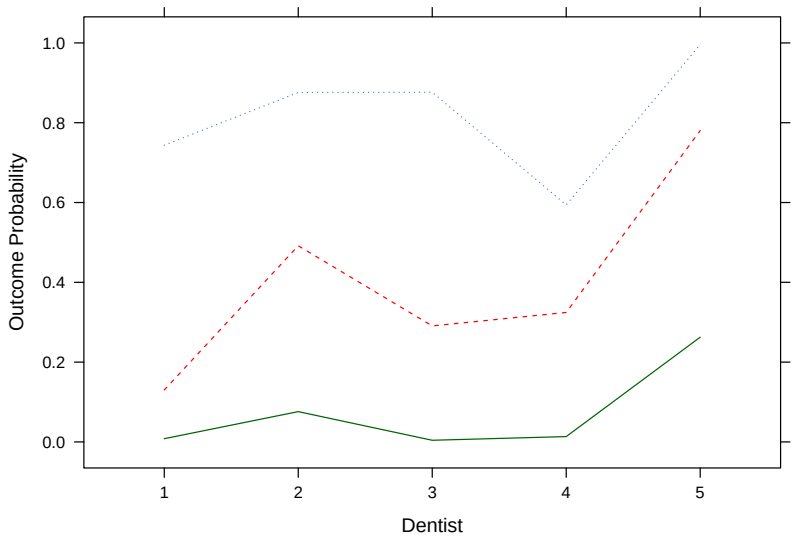

Figure 2: Outcome probabilities for 3 Class Latent Class model.

```
> trellis.par.set(col.whitebg())
> print(plot(dentistry.lca2,type="l",xlab="Dentist",
+     ylab="Outcome Probability"))
```

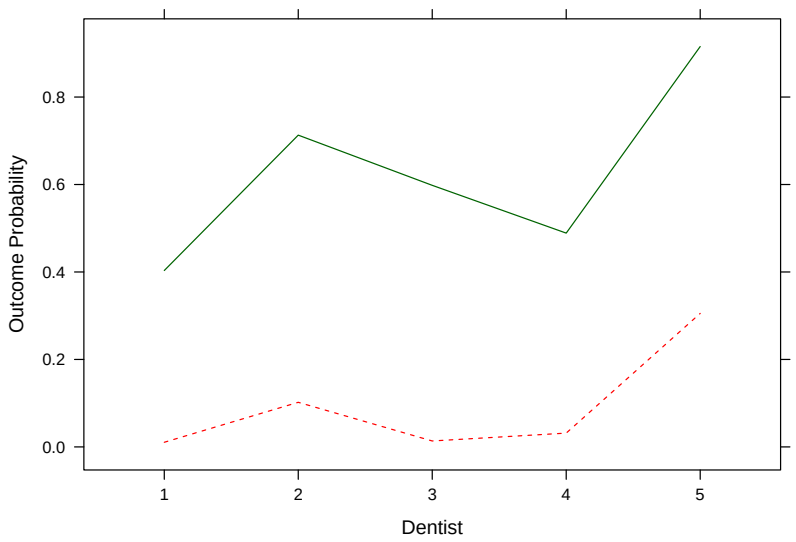

Figure 3: Outcome probabilities for 2 Class Latent Class model.

which is assumed to, sometiems optimistically, correctly classify the subject. The latent class method constructs a hypothetical standard, which has the disadvantage that this is not known with certainty but it allows correctly for any uncertainty in the underlying disease state. The other measure is specificity which is the probability of correctly identifying a subject as not diseased. The sensitivities are then simply the outcome probabilities for the diseased class, Class 1 and the specificity one minus the outcome probabilities for the non-diseased class, Class 2. These can be obtained with 95% confidence intervals using the outcome.probs function.

```
> outcome.probs(dentistry.lca2)

Class  1
   Outcome p     2.5 %    97.5 %
V1 0.4033507 0.3616410 0.4465062
V2 0.7128814 0.6691323 0.7529812
V3 0.5981282 0.5494260 0.6449679
V4 0.4888447 0.4468917 0.5309554
V5 0.9154706 0.8856534 0.9380563
Class  2
    Outcome p       2.5 %     97.5 %
V1 0.01061848 0.006938933 0.01621735
V2 0.10198787 0.089733658 0.11570283
V3 0.01359122 0.008592972 0.02143394
V4 0.03156300 0.024211228 0.04105323
V5 0.30527870 0.287118959 0.32406489
```

The sensitivity and specificity are shown in Table 5. A reasonable conclusion is that the dentists are fairly good at identifying teeth that are not diseased (except for dentist 5), but not too good at identifying teeth that are diseased.

```
> probs <- outcome.probs(dentistry.lca2)
> # this swaps around the probabilities based on the knowledge that
> # the outcome probabilities are higher in the diseased class
> order <- ifelse(probs[[1]][1,1]<probs[[2]][1,1],2,1)
> spec <- NULL
> sens <- NULL
> for (i in 1:5) {
+     sens <- c(sens,sprintf("%3.2f (%3.2f,%3.2f)",probs[[order]]$Outcome[i],
+         probs[[order]]$"2.5 %"[i],probs[[order]]$"97.5 %"[i]))
```

```
+     spec <- c(spec,sprintf("%3.2f (%3.2f,%3.2f)",1-probs[[3-order]]$Outcome[i
+         1-probs[[3-order]]$"97.5 %"[i],1-probs[[3-order]]$"2.5 %"[i]))
+ }
> stable <- data.frame(sens,spec)
> names(stable) <- c("Sensitivity","Specificity")
> row.names(stable) <- paste("V",1:5,sep="")

> print(xtable(stable, digits = c(0,2,2),
+         caption="Sensitivity and Specificity",
+         label="tab:outcomeconfint"),include.rownames=TRUE)
```

	Sensitivity	Specificity
V1	0.40 (0.36,0.45)	0.99 (0.98,0.99)
V2	0.71 (0.67,0.75)	0.90 (0.88,0.91)
V3	0.60 (0.55,0.64)	0.99 (0.98,0.99)
V4	0.49 (0.45,0.53)	0.97 (0.96,0.98)
V5	0.92 (0.89,0.94)	0.69 (0.68,0.71)

Table 5: Sensitivity and Specificity

The confidence intervals for the outcome probabilities can be calculated using the parametric bootstrap. These are shown in Table 6 and are in agreement with those from the standard errors.

	Sensitivity	Specificity
V1	0.40 (0.36,0.44)	0.99 (0.98,0.99)
V2	0.71 (0.67,0.75)	0.90 (0.88,0.91)
V3	0.60 (0.55,0.65)	0.99 (0.98,0.99)
V4	0.49 (0.45,0.53)	0.97 (0.96,0.97)
V5	0.92 (0.89,0.94)	0.69 (0.68,0.71)

Table 6: Sensitivity and Specificity

The true and false positive rates can be plotted for each dentist, and are shown in Figure 4. This gives a better explanation of what is happening. It appears that the difference between dentists is mainly related to the threshold for what they classify as diseased. Dentist 5 is more likely to correctly identify teeth as diseased but at the expense of being more likely to identify non-diseased teeth as diseased.

```
> itpr <- ifelse(dentistry.lca2$classp[2]>dentistry.lca2$classp[1],1,2)
> ifpr <- 3-itpr
```

```
> trellis.par.set(col.whitebg())
> print(plot(tpr~fpr,type="p",
+         xlab="False Positive Rate\n(1-Specificity)",
+         ylab="True Positive Rate (Sensitivity)",data=probs))

NULL
```

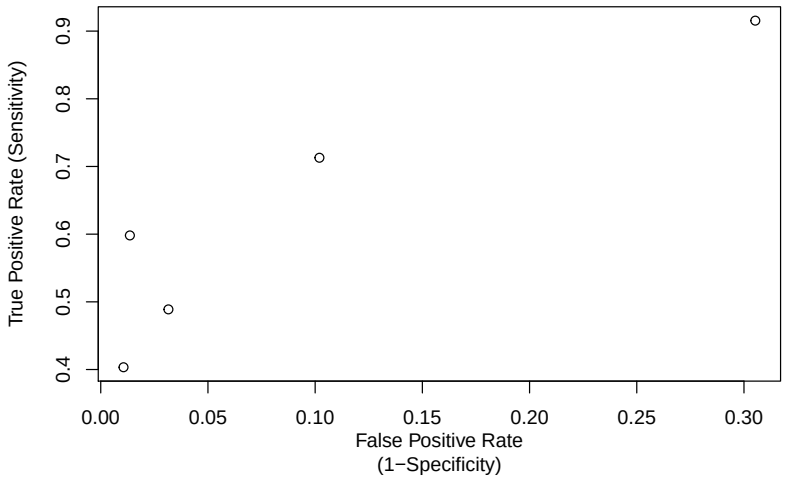

Figure 4: True and False Positive Rates by Dentist.

```
> probs <- outcome.probs(dentistry.lca2)
> probs <- data.frame(tpr=probs[[itpr]][,1],fpr=probs[[ifpr]][,1])
```

3 Latent Class with Random Effects

3.1 Model

The method used in Qu et al. (1996) is to introduce a random effect to model heterogeneity within classes. In their model the probabilities are transformed to the probit scale and then a normal random effect introduced. In practice it usually makes little difference if a probit or logit transform is used.

The probability for each observation remains the same, except that it is now conditional on both class c and random effect λ.

$$P\left(y_{i1}, y_{i2}, ..., y_{ik}|c, u\right) = \prod_{j=1}^{k} \pi_{cj}^{y_{ij}} \left(1 - \pi_{cj}\right)^{1-y_j}$$

Where

$$\pi_{cj} = \frac{e^{a_{cj}+b_j u}}{1 + e^{a_{cj}+b_j u}}, u \sim N(0, 1)$$

b_j scales the random effect - models may have either a common or independent scale for each outcome, these are the `lambdacoef`. They may also be chosen to be different for each class, the default is for them to be the same for each class.

One way of visualising the model is that each class is now an Item Response Theory (IRT) model when the scaling is independent. When the scaling is common, the loadings are the same for each outcome and each class is then a Rasch model.

3.2 Example 2 Continued

We now continue the analysis of the dentistry data, allowing for random effects. This has a simple interpretation. For each subject there will be a different level of disease, and as a result a dentist will be more or less likely to classify the subject as having disease.

```
> dentistry.lca2random <- randomLCA(dentistry[,1:5],freq=dentistry$freq,
+         nclass=2,random=TRUE,quadpoints=41,probit=TRUE)
```

The BIC is reduced to 14944.7 showing an improvement over any of the latent class models. An alternative model is to allow the variance of the random effect to vary by outcome (dentist). This can be performed using the `blocksize` parameter. This allows the structure of the data to be set as a series of blocks, and within each block each outcome has a different loading.

```
> dentistry.lca2random1 <- randomLCA(dentistry[,1:5],freq=dentistry$freq,
+         nclass=2,random=TRUE,probit=TRUE,
+         quadpoints=41,blocksize=5)
```

```
> trellis.par.set(col.whitebg())
> print(plot(dentistry.lca2random1,graphtype="marginal",type="l",xlab="Dentist"
+     ylab="Marginal Outcome Probability"))

NULL
```

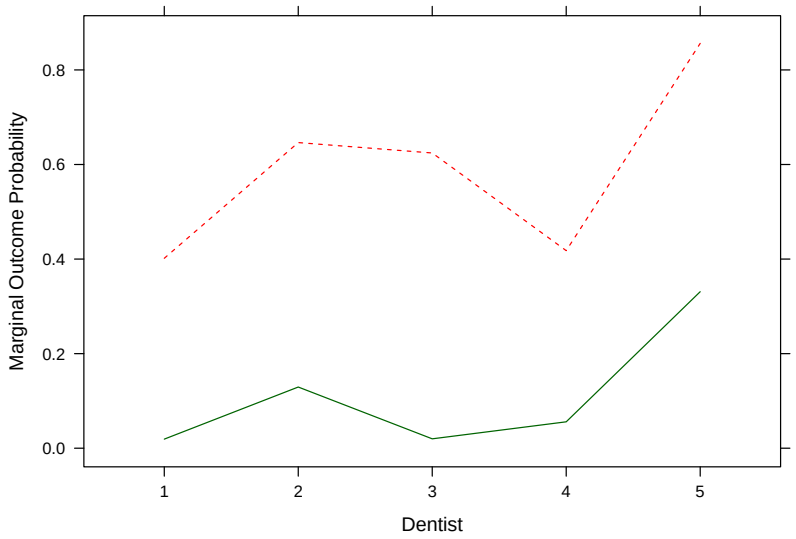

Figure 5: Marginal Outcome Probabilities for 2 Class Latent Class with Random Effect (2LCR) model.

This increases the BIC to 14944.7, and is the 2LCR model obtained by Qu et al. (1996). It appears that the simpler model is more appropriate.

A further extension is to allow the loading or random effect variance to vary by class.

```
> dentistry.lca2random2 <- randomLCA(dentistry[,1:5],freq=dentistry$freq,
+         nclass=2,random=TRUE,probit=TRUE,
+         blocksize=5,byclass=TRUE,quadpoints=41)
```

The BIC increases to 14951.1. It is not surprising that this model isn't an improvement, there are now 21 parameters fitted to 32 observations. This also may give problems with the fitting algorithm so the number of quadrature points is increased to 41.

The marginal outcome probabilities, obtained by integrating over the random effect can be plotted, as in Figure 5.

Outcome probabilities with confidence intervals may be calculated for models with random effects only using the parametric bootstrap. The sensitivity and specificity may be obtained from these and are shown in Table 7. Differences from the Qu et al paper result from them using a model with an individual loading for each dentist when calculating their Table 6.

```
> print(xtable(stable, digits = c(0,2,2),
+         caption="Sensitivity and Specificity",
+         label="tab:outcomeconfintboot2"),include.rownames=TRUE)
```

	Sensitivity	Specificity
V1	0.40 (0.34,0.48)	0.98 (0.97,0.99)
V2	0.65 (0.56,0.73)	0.87 (0.85,0.89)
V3	0.62 (0.51,0.73)	0.98 (0.73,1.00)
V4	0.42 (0.35,0.49)	0.94 (0.93,0.96)
V5	0.86 (0.77,0.91)	0.67 (0.64,0.69)

Table 7: Sensitivity and Specificity

The observed and fitted values can be obtained and are shown in Table 8. Differences from the Qu et al paper again result from their different model.

```
> obs.data <- data.frame(dentistry.lca2random1$patterns,dentistry.lca2random1$o
+         dentistry.lca2$fitted,dentistry.lca2random1$fitted)
> names(obs.data) <- c("V1","V2","V3","V4","V5","Obs","Exp 2LC","Exp 2LCR")
```

V1	V2	V3	V4	V5	Obs	Exp 2LC	Exp 2LCR
0	0	0	0	0	1880	1836.3	1882.2
0	0	0	0	1	789	830.4	779.8
0	0	0	1	0	43	61.9	56.1
0	0	0	1	1	75	49.6	72.3
0	0	1	0	0	23	28.6	25.8
0	0	1	0	1	63	47.5	60.4
0	0	1	1	0	8	4.0	4.7
0	0	1	1	1	22	35.1	25.1
0	1	0	0	0	188	213.9	176.1
0	1	0	0	1	191	152.2	209.7
0	1	0	1	0	17	12.1	17.6
0	1	0	1	1	67	61.0	53.8
0	1	1	0	0	15	11.2	14.5
0	1	1	0	1	85	91.6	79.0
0	1	1	1	0	8	8.1	5.6
0	1	1	1	1	56	86.4	67.1
1	0	0	0	0	22	21.2	17.0
1	0	0	0	1	26	25.2	30.6
1	0	0	1	0	6	2.1	2.5
1	0	0	1	1	14	16.1	10.6
1	0	1	0	0	1	2.5	3.3
1	0	1	0	1	20	24.7	20.8
1	0	1	1	0	2	2.2	1.4
1	0	1	1	1	17	23.5	17.8
1	1	0	0	0	2	6.0	7.4
1	1	0	0	1	20	42.0	31.9
1	1	0	1	0	6	3.7	2.4
1	1	0	1	1	27	39.3	23.6
1	1	1	0	0	3	5.7	4.6
1	1	1	0	1	72	61.1	59.4
1	1	1	1	0	1	5.4	3.5
1	1	1	1	1	100	58.4	102.5

Table 8: Observed and expected frequencies

References

Y Qu, M Tan, and MH Kutner. Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics*, 52(3):797–810, 1996.

D Rindskopf and W Rindskopf. The value of latent class analysis in medical diagnosis. *Statistics in Medicine*, 5:21–27, 1986.