randomLCA Examples

Ken Beath

March 8, 2009

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 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, calcSE = TRUE)
> 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
        2 379.3954 402.2851 -180.6977
Class probabilities
Class  1 Class  2
  0.5422   0.4578
Outcome probabilities
            Q.wave History     LDH    CPK
Class  1 1.964e-08  0.1951 0.02692 0.1955
Class  2 7.668e-01  0.7914 0.82791 1.0000
```

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	0.000	0.195	0.027	0.196
Class 2	0.767	0.791	0.828	1.000

Table 2: Outcome Probabilities.

This gives some interesting information. In Class 1, 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 0.54 and 0.46 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 class probs can be obtained as

```
> classprobs <- cbind(myocardial.lca2$patterns,
+     myocardial.lca2$classprob)
> colnames(classprobs) <- c(names(myocardial)[1:4],
+     "Class 1", "Class 2")
```

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.

Outcome probabilities are shown in Figure 1.

3 Example 2

This example shows the fitting of the dentistry data from Qu, Tan and Kutner (1996). The data consists of the results of five dentists evaluating x-rays for

Q.wave	History	LDH	CPK	Class 1	Class 2
1	1	1	1	0.000	1.000
0	1	1	1	0.008	0.992
1	0	1	1	0.000	1.000
0	0	1	1	0.111	0.889
1	1	0	1	0.000	1.000
0	1	0	1	0.581	0.419
1	0	0	1	0.000	1.000
0	0	0	1	0.956	0.044
1	1	1	0	0.662	0.338
0	1	1	0	1.000	0.000
1	0	1	0	0.968	0.032
0	0	1	0	1.000	0.000
1	1	0	0	0.997	0.003
0	1	0	0	1.000	0.000
1	0	0	0	1.000	0.000
0	0	0	0	1.000	0.000

Table 3: Class Probabilities.

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.

3.1 Latent Class

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

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

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)))
```

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

```
> 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]))))
```

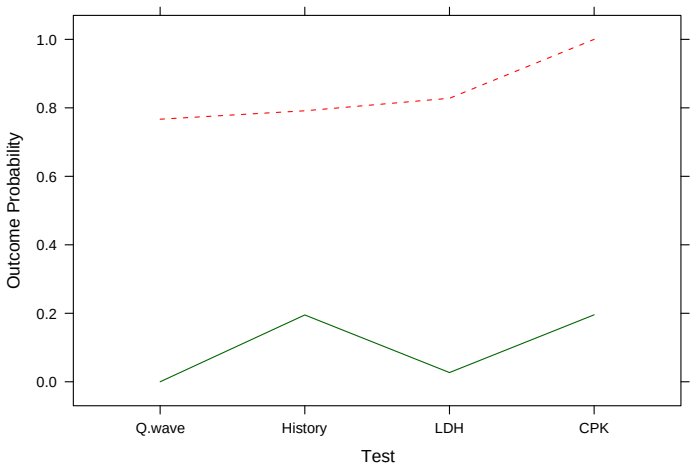

Figure 1: Outcome probabilities for 2 Class Latent Class model.

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

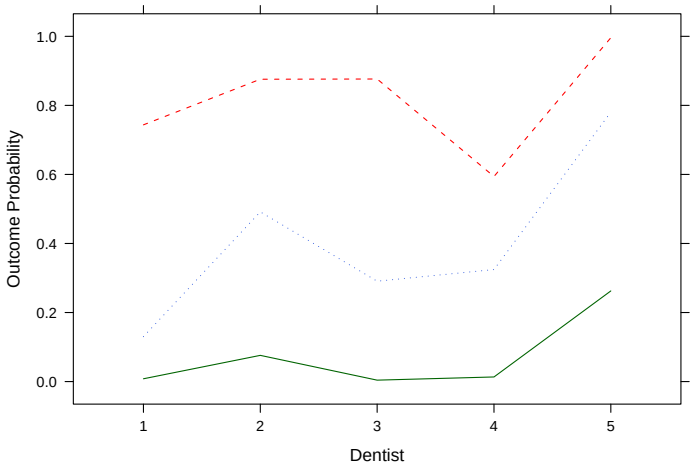

Figure 2: Outcome probabilities for 3 Class Latent Class model.

classes	bic
1	17531.1
2	15021.6
3	14962.9
4	15000.0

Table 4: BIC by class.

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

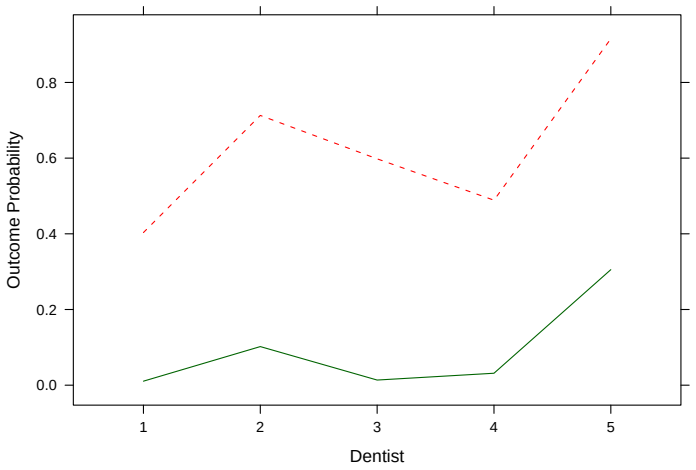

Figure 3: Outcome probabilities for 2 Class Latent Class model.

given that the subject is diseased. In classical diagnostic testing the "true" status of a subject is known through use of a "gold standard" 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 2 and the specificity one minus the outcome probabilities for the non-diseased class, Class 1. These can be obtained with 95% confidence intervals (provided the model is fitted with calcSE=TRUE) using the outcome.probs function.

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

```
Class  1
    Outcome p       2.5 %     97.5 %
V1 0.01061847 0.006938941 0.01621728
V2 0.10198786 0.089733773 0.11570266
```

```
V3 0.01359118 0.008592960 0.02143383
V4 0.03156297 0.024211316 0.04105300
V5 0.30527866 0.287119155 0.32406459
Class  2
   Outcome p     2.5 %    97.5 %
V1 0.4033506 0.3616416 0.4465053
V2 0.7128811 0.6691326 0.7529805
V3 0.5981282 0.5494269 0.6449670
V4 0.4888446 0.4468918 0.5309552
V5 0.9154705 0.8856535 0.9380562
```

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)
> order <- ifelse(dentistry.lca2$classp[2] > dentistry.lca2$classp[1],
+     1, 2)
> 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]]$"2.5 %"[i], 1 -
+             probs[[3 - order]]$"97.5 %"[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 (0.99,0.98)
V2	0.71 (0.67,0.75)	0.90 (0.91,0.88)
V3	0.60 (0.55,0.64)	0.99 (0.99,0.98)
V4	0.49 (0.45,0.53)	0.97 (0.98,0.96)
V5	0.92 (0.89,0.94)	0.69 (0.71,0.68)

Table 5: 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

```
> 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
```

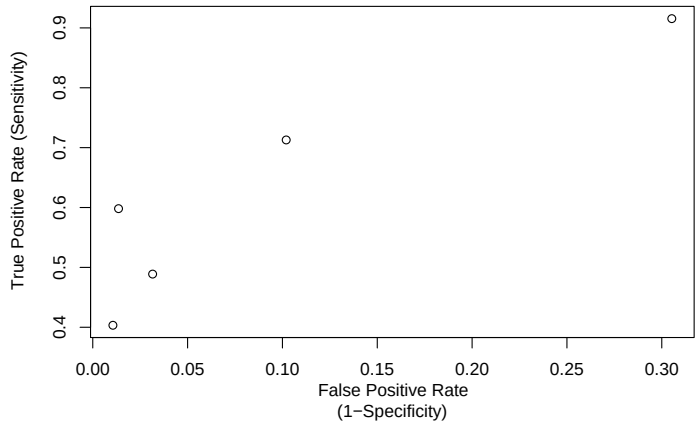

Figure 4: True and False Positive Rates by Dentist.

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

3.2 Latent Class with Random Effects

The method used in Qu, Tan and Kutner (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.

```
> dentistry.lca2random <- randomLCA(dentistry[,
+     1:5], freq = dentistry$freq, initmodel = dentistry.lca2,
+     nclass = 2, random = TRUE, 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, initmodel = dentistry.lca2random,
+     nclass = 2, random = TRUE, probit = TRUE,
+     blocksize = 5)
```

This increases the BIC to 14949.4, and is the 2LCR model obtained by Qu, Tan and Kutner (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, initmodel = dentistry.lca2random1,
+     nclass = 2, random = TRUE, probit = TRUE,
+     blocksize = 5, byclass = TRUE, quadpoints = 41)
```

The BIC increases to 14987.6. 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 increase to 41.

The observed and fitted values can be obtained and are shown in Table 6. Differences from the Que et al paper result from different approximation methods.

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

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

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

NULL
```

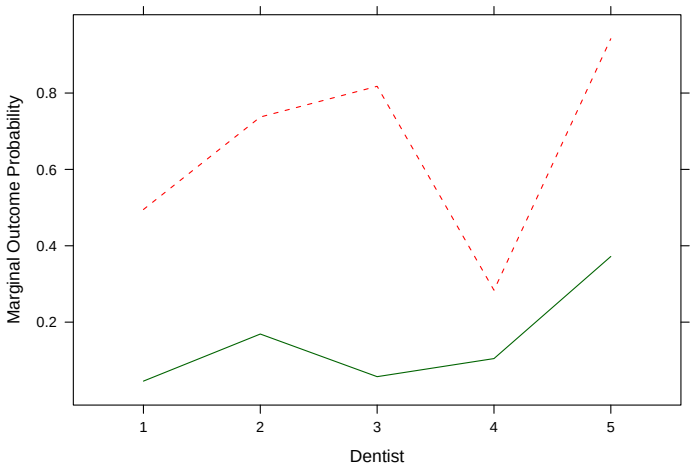

Figure 5: Marginal Outcome Probabilities for 2 Class Latent Class with Random Effect (2LCR) model.

V1	V2	V3	V4	V5	Obs	Exp 2LC	Exp 2LCR
0	0	0	0	0	1880	1836.3	1882.6
0	0	0	0	1	789	830.4	784.7
0	0	0	1	0	43	61.9	38.2
0	0	0	1	1	75	49.6	79.7
0	0	1	0	0	23	28.6	24.2
0	0	1	0	1	63	47.5	63.8
0	0	1	1	0	8	4.0	6.8
0	0	1	1	1	22	35.1	25.8
0	1	0	0	0	188	213.9	184.7
0	1	0	0	1	191	152.2	192.5
0	1	0	1	0	17	12.1	23.1
0	1	0	1	1	67	61.0	67.2
0	1	1	0	0	15	11.2	12.5
0	1	1	0	1	85	91.6	87.4
0	1	1	1	0	8	8.1	7.1
0	1	1	1	1	56	86.4	50.8
1	0	0	0	0	22	21.2	18.5
1	0	0	0	1	26	25.2	27.9
1	0	0	1	0	6	2.1	4.8
1	0	0	1	1	14	16.1	16.0
1	0	1	0	0	1	2.5	2.3
1	0	1	0	1	20	24.7	19.7
1	0	1	1	0	2	2.2	1.8
1	0	1	1	1	17	23.5	14.5
1	1	0	0	0	2	6.0	7.3
1	1	0	0	1	20	42.0	19.8
1	1	0	1	0	6	3.7	4.7
1	1	0	1	1	27	39.3	22.4
1	1	1	0	0	3	5.7	2.7
1	1	1	0	1	72	61.1	69.6
1	1	1	1	0	1	5.4	3.2
1	1	1	1	1	100	58.4	103.0

Table 6: Observed and expected frequencies