Latent class analysis

Classification and covariates

DL Oberski & L Boeschoten

Classification

(Putting people into boxes, while admitting uncertainty)

Classification

- After estimating a LC model, we may wish to classify individuals into latent classes

- The latent classification or **posterior** class membership probabilities $P(X = x \mid \mathbf{y})$ can be obtained from the LC model parameters using Bayes' rule:

$$P(X = x \mid \mathbf{y}) = \frac{P(X = x)P(\mathbf{y} \mid X = x)}{P(\mathbf{y})} = \frac{P(X = x)\prod_{k=1}^{K} P(y_k \mid X = x)}{\sum_{c=1}^{C} P(X = c)\prod_{k=1}^{K} P(y_k \mid X = c)}$$

Small example: posterior classification

Y1	Y2	Y3	P(X=1 | Y)	P(X=2 | Y)	Most likely (but not sure!)
1	1	1	0.002	0.998	2
1	1	2	0.071	0.929	2
1	2	1	0.124	0.876	2
1	2	2	0.832	0.169	1
2	1	1	0.152	0.848	2
2	1	2	0.862	0.138	1
2	2	1	0.920	0.080	1
2	2	2	0.998	0.003	1

Classification quality

Classification Statistics

- classification table: true vs. assigned class
- overall proportion of classification errors

Other reduction of "prediction" errors measures

- How much more do we know about latent class membership after seeing the responses?
- Comparison of $P(X=x)$ with $P(X=x \mid \mathbf{Y}=\mathbf{y})$
- R-squared-like reduction of prediction (of X) error

```r
posteriors <- data.frame(M4$posterior, predclass=M4$predclass)

classification_table <-
    ddply(posteriors, .(predclass), function(x) colSums(x[,1:4])))

> round(classification_table, 1)
 predclass post.1 post.2 post.3 post.4
1         1 1824.0   34.9    0.0   11.1
2         2    7.5   87.4    1.1    3.0
3         3    0.0    1.0   19.8    0.2
4         4    4.0    8.6    1.4   60.1
```

Classification table for 4-class

	post.1	post.2	post.3	post.4
1	**0.99**	0.26	0.00	0.15
2	0.00	**0.66**	0.05	0.04
3	0.00	0.01	**0.89**	0.00
4	0.00	0.07	0.06	**0.81**
	1	1	1	**1**

Total classification errors:

```
> 1 - sum(diag(classification_table)) / sum(classification_table)
[1] 0.0352
```

Entropy R^2

```
entropy <- function(p) sum(-p * log(p))
error_prior <- entropy(M4$P) # Class proportions
error_post <- mean(apply(M4$posterior, 1, entropy))

R2_entropy  <- (error_prior - error_post) / error_prior

> R2_entropy
[1] 0.741
```

This means that we know a lot more about people's political participation class after they answer the questionnaire.

Compared with if we only knew the overall proportions of people in each class

Classify-analyze can give some bias

- You might think that after classification it is easy to model people's latent class membership

- "Just take assigned class and analyze it as though observed"

- Unfortunately, -> **biased estimates** and wrong se's

  (Still relatively little-known among practicioners?)

Predicting latent class membership

(using covariates; concomitant variables)

Fitting a LCM in poLCA with gender as a covariate

```
M4 <- poLCA(
     cbind(contplt, wrkprty, wrkorg,
          badge, sgnptit, pbldmn, bctprd) ~ gndr,
          data=gr, nclass = 4, nrep=20)
```

This gives a **multinomial logistic regression** with X as dependent and gender as independent (“concomitant”; “covariate”)

Predicting latent class membership from a covariate

$$P(X=x \mid Z=z)=\frac{\exp(\gamma_{0x}+\gamma_{zx})}{\sum_{c=1}^{C}\exp(\gamma_{0c}+\gamma_{zc})}$$

γ_{0x} Is the logistic intercept for category x of the latent class variable X

γ_{zx} Is the logistic slope predicting membership of class x for value z of the covariate Z

```
=========================================================
Fit for 4 latent classes:
=========================================================
2 / 1
            Coefficient  Std. error  t value  Pr(>|t|)
(Intercept)    -0.35987     0.37146   -0.969     0.335
gndrFemale     -0.34060     0.39823   -0.855     0.395
=========================================================
3 / 1
            Coefficient  Std. error  t value  Pr(>|t|)
(Intercept)     2.53665     0.21894   11.586     0.000
gndrFemale      0.21731     0.24789    0.877     0.383
=========================================================
4 / 1
            Coefficient  Std. error  t value  Pr(>|t|)
(Intercept)    -1.57293     0.39237   -4.009     0.000
gndrFemale     -0.42065     0.57341   -0.734     0.465
=========================================================
```

One possible interpretation

Class 1 Modern political participation

Class 2 Traditional political participation

Class 3 No political participation

Class 4 Every kind of political participation

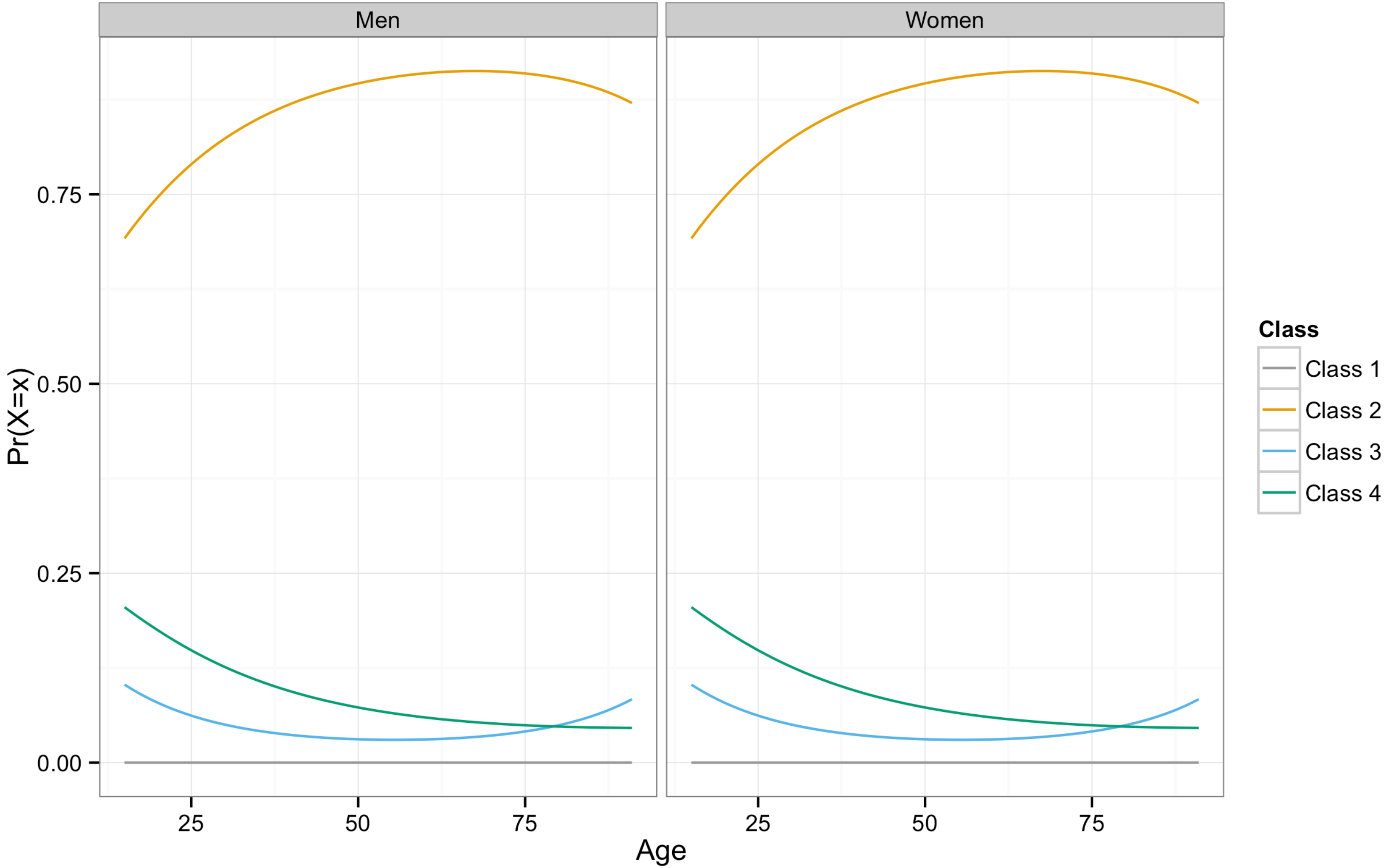
Women more likely than men to be in classes 1 and 3

Less likely to be in classes 2 and 4

Multinomial logistic regression refresher

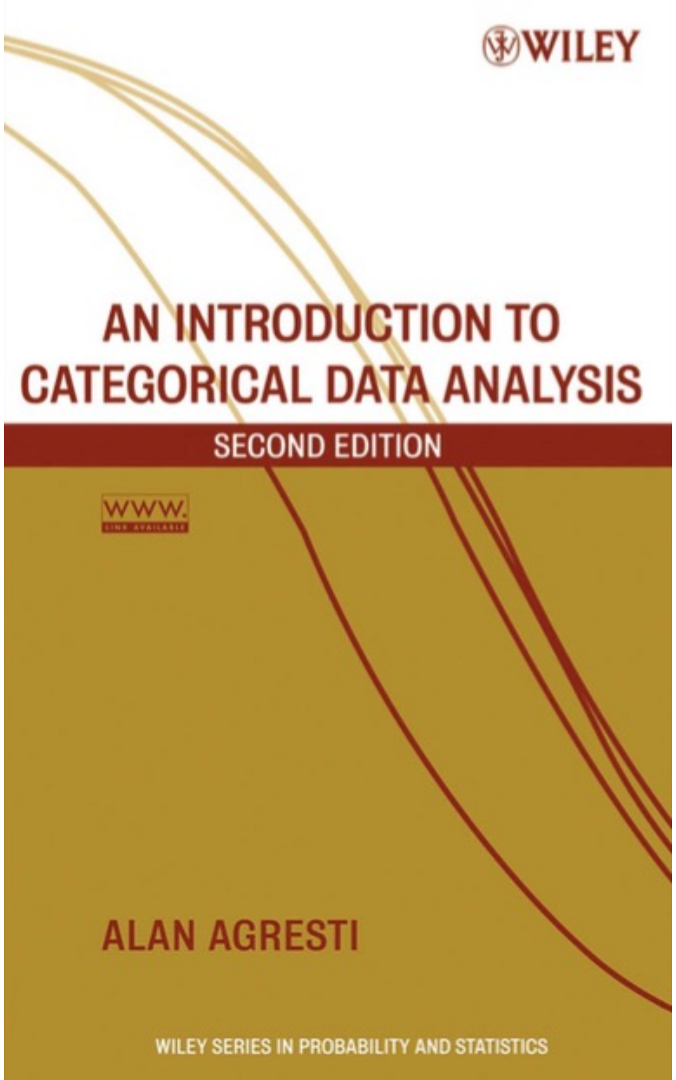
For example:

- Logistic multinomial regression coefficient equals -0.3406
- Then log odds ratio of being in class 2 (compared with reference class 1) is -0.3406 smaller for women than for men
- So odds ratio is smaller by a factor exp(-0.3406) = 0.71
- So odds are 30% smaller for women



Men
Women
Pr(X=x)
0.75
0.50
0.25
0.00
25
50
75
Age
Class
Class 1
Class 2
Class 3
Class 4

Even more (re)freshing:



Problems you will encounter when doing latent class analysis

(and some solutions)

Some problems

- Local maxima
- Boundary solutions
- Non-identification

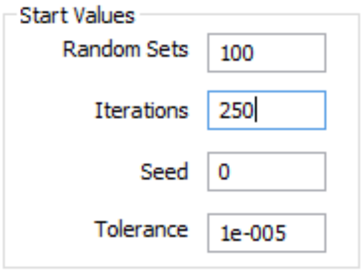
Problem: Local maxima

Problem: there may be different sets of "ML" parameter estimates with different L-squared values we want the solution with lowest L-squared (highest log-likelihood)

Solution: multiple sets of starting values

```
poLCA(cbind(Y1, Y2, Y3)~1, antireli, nclass=2, nrep=100)


Model 1: llik = -3199.02 ... best llik = -3199.02
Model 2: llik = -3359.311 ... best llik = -3199.02
Model 3: llik = -2847.671 ... best llik = -2847.671
Model 4: llik = -2775.077 ... best llik = -2775.077
Model 5: llik = -2810.694 ... best llik = -2775.077
....
```



Problem: boundary solutions

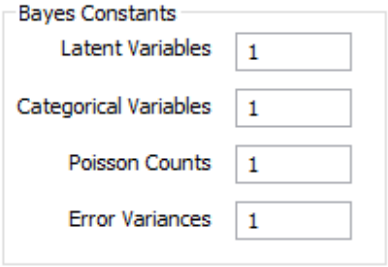
Problem: estimated probability becomes zero/one, or logit parameters extremely large negative/positive

Example:

```
$badge
           Pr(1)  Pr(2)
class 1:  0.8640 0.1360
class 2:  0.1021 0.8979
class 3:  0.4204 0.5796
class 4:  0.0000 1.0000
```

Solutions:

1. Not really a problem, just ignore it;
2. Use priors to smooth the estimates
3. Fix the offending probabilities to zero (classical)



Problem: non-identification

- Different sets of parameter estimates yield the same value of L-squared and LL value: estimates are not unique
- Necessary condition DF>=0, but not sufficient
- Detection: running the model with different sets of starting values or, formally, checking whether rank of the Jacobian matrix equals the number of free parameters
- “Well-known” example: 3-cluster model for 4 dichotomous indicators

What we did not cover

- 1 step versus 3 step modeling
- Ordinal, continuous, mixed type indicators
- Hidden Markov (“latent transition”) models
- Mixture regression

What we did cover

- Latent class “cluster” analysis
- Model formulation, different parameterizations
- Model interpretation, profile
- Model fit evaluation: global, local, and substantive
- Classification
- Common problems with LCM and their solutions