Solutions Undirected Graphical Models

Exercise 1

(a) $N = 2026$, $n(\text{female}, \text{brown}) = 352$, and $n(\text{hazel}) = 347$.

(b) For example

$$\hat{n}(\text{male}, \text{green}) = \frac{n(\text{male})n(\text{green})}{N} = \frac{919 \times 308}{2026} = 139.71$$

The other cells in the table of fitted counts are computed in a similar way. This yields:

	Eye Color				
Gender	blue	brown	green	hazel	Total
female	398.32	350.79	168.29	189.60	1107
male	330.68	291.21	139.71	157.40	919
Total	729	642	308	347	2026

(c) The cliques of the independence graph are the individual nodes of gender and eye color, so we have the margin constraints:

$$\hat{n}(\text{gender}) = n(\text{gender})$$
$$\hat{n}(\text{eye color}) = n(\text{eye color})$$

The IPF algorithm fits the counts to each margin in turn, and repeats this process until all margin constraints are satisfied simultaneously. For the algorithm to work correctly, we should start from a solution that satisfies all constraints of the model to be fitted: if the model puts a u-term to zero, it should also have the value 0 in our initial solution $\hat{n}^{(0)}$. Therefore, starting from the uniform table is a safe choice, because it puts all u-terms to zero except $u_\emptyset$. Which particular count we put in all cells is not important. So take $\hat{n}^{(0)}$ to be:

	Eye Color				
Gender	blue	brown	green	hazel	Total
female	1	1	1	1	4
male	1	1	1	1	4
Total	2	2	2	2	8

To obtain $\hat{n}^{(1)}$, we fit to the observed row margin:

	Eye Color				
Gender	blue	brown	green	hazel	Total
female					1107
male					919
Total					2026

We distribute the row total over the columns according to $\hat{P}^{(0)}(\text{Eye Color}|\text{Gender})$, so for example

$$\hat{P}^{(0)}(\text{blue}|\text{female}) = \frac{1}{4},$$

so the cell (female,blue) gets a fitted count of $\hat{n}^{(1)}(\text{female}, \text{blue}) = 1107 \times \frac{1}{4} = 276.75$. Completing the table in this way, $\hat{n}^{(1)}$ becomes:

	Eye Color				
Gender	blue	brown	green	hazel	Total
female	276.75	276.75	276.75	276.75	1107
male	229.75	229.75	229.75	229.75	919
Total	506.5	506.5	506.5	506.5	2026

Now the row margin is correct, but the column margin is off. To obtain $\hat{n}^{(2)}$, we fit to the observed column margin:

	Eye Color				
Gender	blue	brown	green	hazel	Total
female					
male					
Total	729	642	308	347	2026

We distribute the column total over the rows according to $\hat{P}^{(1)}(\text{Gender}|\text{Eye Color})$, so for example

$$\hat{P}^{(1)}(\text{female}|\text{blue}) = \frac{276.75}{506.5} = 0.5463986$$

so the cell (female,blue) gets a fitted count of $\hat{n}^{(2)}(\text{female}, \text{blue}) = 729 \times 0.5463986 = 398.32$. Completing the table in this way, $\hat{n}^{(2)}$ becomes:

	Eye Color				
Gender	blue	brown	green	hazel	Total
female	398.32	350.79	168.29	189.60	1107
male	330.68	291.21	139.71	157.40	919
Total	729	642	308	347	2026

Now both margin constraints are satisfied simultaneously, so the algorithm has converged. As a general rule, if closed form estimates exist (the model is decomposable), then the IPF algorithm converges in one cycle through all margins that have to be fitted.

(d) I don't know!

(e) We test the independence model against the saturated model. The degrees of freedom for the χ^2 test is equal to the difference in the number of u-terms of the two models. The log-linear expansion of the saturated model is:

$$\log P(\text{gender}, \text{eye color}) = u_\emptyset + u(\text{gender}) + u(\text{eye color}) + u(\text{gender}, \text{eye color})$$

The log-linear expansion for the independence model is:

$$\log P(\text{gender}, \text{eye color}) = u_\emptyset + u(\text{gender}) + u(\text{eye color})$$

The independence model excludes all u-terms $u(\text{gender}, \text{eye color})$. How many are there? Number the values of gender as 0 and 1, and number the values of eye color as 0,1,2,3. If either variable has the value 0, then $u(\text{gender}, \text{eye color}) = 0$. So the number of non-zero such u-terms is $1 \times 3 = 3$. In the table we look up $\chi^2_{3;0.05} = 7.82$. The observed deviance is 16.29, which is bigger than the critical value of 7.82, so we reject the null hypothesis that the independence model is the true model.

In general, if we have an $r \times c$ table (where r is the number of rows and c the number of columns) and we test the independence model against the saturated model, then the appropriate degrees of freedom for the test is $(r-1) \times (c-1)$.

(f) The cliques of the independence graph are the individual nodes of gender and eye color, so we have the margin constraints:

$$\hat{n}(\text{gender}) = n(\text{gender})$$
$$\hat{n}(\text{eye color}) = n(\text{eye color})$$

Because of the independence assumption we have

$$\hat{P}(\text{gender}, \text{eye color}) = \hat{P}(\text{gender})\hat{P}(\text{eye color})$$

Multiplying on the left by N and on the right by $\frac{N^2}{N} = N$ we get

$$\hat{n}(\text{gender}, \text{eye color}) = \frac{\hat{n}(\text{gender})\hat{n}(\text{eye color})}{N}$$

Finally, we use the margin constraints to obtain

$$\hat{n}(\text{gender}, \text{eye color}) = \frac{n(\text{gender})n(\text{eye color})}{N}$$

Exercise 2

(a) {3} separates {1,2} from {4,5} because every path from a node in {1,2} to a node in {4,5} has to pass through node 3. Therefore we may conclude that the conditional independence

$$(X_1, X_2) \perp\!\!\!\perp (X_4, X_5) | X_3$$

holds.

(b) $P(X_1, X_2, X_4, X_5|X_3) = P(X_1, X_2|X_3)P(X_4, X_5|X_3)$.

(c) The cliques are {1,2,3} and {3,4,5}. The corresponding margin constraints are

$$\hat{n}\,(X_1, X_2, X_3) = n(X_1, X_2, X_3)$$
$$\hat{n}(X_3, X_4, X_5) = n(X_3, X_4, X_5)$$

(d) Make sure you justify each step:

$$\begin{aligned}
\hat{P}(X_1, X_2, X_3, X_4, X_5) &= \hat{P}(X_1, X_2, X_4, X_5|X_3)\hat{P}(X_3) && \text{(product law)} \\
&= \hat{P}(X_1, X_2|X_3)\hat{P}(X_4, X_5|X_3)\hat{P}(X_3) \\
& && ((X_1, X_2) \perp\!\!\!\perp (X_4, X_5) \mid X_3) \\
&= \frac{\hat{P}(X_1, X_2, X_3)\hat{P}(X_3, X_4, X_5)}{\hat{P}(X_3)} && \text{(product law twice)}
\end{aligned}$$

We have reached our goal: in the numerator we have distributions over the cliques, and in the denominator over a subset of a clique. Now we multiply by N on the left and by $N^2/N = N$ on the right to get fitted counts instead of fitted probabilities:

$$\hat{n}(X_1, X_2, X_3, X_4, X_5) = \frac{\hat{n}(X_1, X_2, X_3)\hat{n}(X_3, X_4, X_5)}{\hat{n}(X_3)}$$

Finally, we can use the property that the maximum likelihood solution satisfies the margin constraints (fitted = observed for every margin corresponding to a complete subgraph), so we can replace the fitted counts on the right hand side by observed counts:

$$\hat{n}(X_1, X_2, X_3, X_4, X_5) = \frac{n(X_1, X_2, X_3)n(X_3, X_4, X_5)}{n(X_3)}$$

Excercise 3

(a) There are $\binom{k}{2}$ different edges. Each edge can be either included or excluded, so $2^{\binom{k}{2}}$.

(b) $\binom{8}{2} = 28$. $2^{28} = 268,435,456$. So roughly 268 million.

(c) Graphical: We can remove 7 edges. We can add: AC,AF,BD,BF,CD,CE,CF,DF. That's 8 in total, so there are $7 + 8 = 15$ neighboring graphical models. Could we have found the answer without actually enumerating the possibilities?

Decomposable: We can remove 6 edges (not AE because that would create the chordless 4-cycle $A - B - E - D - A$). We can add every edge, except CF (chordless 4-cylce $B - C - F - E - B$) and CD (chordless 4-cycle $A - B - C - D - A$). So 6+6=12 neighbors.

Exercise 4

(a) P(yes | male) = 245/400=0.6125 and P(yes | female)=75/200=0.375.

(b) Fitted cell counts of the independence model:

Gender	Admission	
	Yes	No
Male	213.33	186.67
Female	106.67	93.33

(c) Value of the deviance:

$$2\left[245\ln\frac{245}{213.33} + 155\ln\frac{155}{186.67} + 75\ln\frac{75}{106.67} + 125\ln\frac{125}{93.33}\right] \approx 30.4$$

(d) The independence model puts one extra u-term to zero compared to the saturated model, so we should use a χ^2 distribution with one degree of freedom. The critical value is

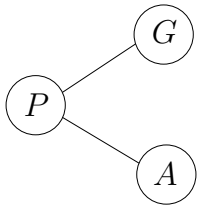
$$\chi^2_{1;0.05} = 3.84.$$

We reject the independence model because the observed deviance is bigger than the critical value.

(e) Clearly, women are less likely to be admitted than men. In itself this does not prove discrimination however. Men and women might differ on other attributes that are legitimate admittance criteria, but that were not taken into account in this analysis (see also the next exercise).

Exercise 5

(a) The independence graph is



Within each program, Gender and Admission are independent.

(b) Maximum likelihood fitted counts:

$$\hat{n}(P, G, A) = \frac{n(P, G)n(P, A)}{n(P)}$$

The fitted counts are:

Program	Gender	Admission	
		Yes	No
A	Male	24.71	80.29
	Female	35.29	114.71
B	Male	222.32	72.68
	Female	37.68	12.32

The deviance is 0.712. Since $\chi^2_{2;0.05} = 6.00$, we don't reject the model.

(c) No. Within program A, the fraction of male applicants that is accepted is $25/105 = 0.24$ and the fraction of female applicants that is accepted is $35/150 = 0.23$, so slightly smaller. However, in program B this is the other way around: 75% of the males is accepted, and 80% of the females.

More women apply to program A, and program A accepts fewer students. That there is no discrimination is confirmed by the good fit of the model $G \perp\!\!\!\perp A|P$.