Topics: Graphical models Due on October 31, 2019

Name of student: Subhrajyoty Roy
Roll number: MB1911

1. Various properties of conditional independence [10]

Prove the following conditional independence statements (for (e) and (f), assume positive pmf/density):

(a) $X \perp\!\!\!\perp Y \mid Z \implies Y \perp\!\!\!\perp X \mid Z$.

(b) $X \perp\!\!\!\perp Y \mid Z$, and $U = h(X) \implies U \perp\!\!\!\perp Y \mid Z$.

(c) $X \perp\!\!\!\perp Y \mid Z$, and $U = h(X) \implies X \perp\!\!\!\perp Y \mid (Z, U)$.

(d) $X \perp\!\!\!\perp Y \mid Z$, and $X \perp\!\!\!\perp W \mid (Y, Z) \implies X \perp\!\!\!\perp (W, Y) \mid Z$.

(e) $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp Z \mid Y \implies X \perp\!\!\!\perp (Y, Z)$.

(f) $X \perp\!\!\!\perp Y \mid (W, Z)$ and $X \perp\!\!\!\perp W \mid (Y, Z) \implies X \perp\!\!\!\perp (W, Y) \mid Z$.

Show that (e) and (f) are false without the assumption of positive pmf/density.

Solution. (a) Since $X \perp\!\!\!\perp Y \mid Z$, we have;

$$f(x, y, z)k(z) = g(x, z)h(y, z)$$

where $f(\cdot), k(\cdot), g(\cdot), h(\cdot)$ represents the respective density. Clearly,

$$f(x, y, z)k(z) = h(y, z)g(x, z)$$

hence, $Y \perp\!\!\!\perp X \mid Z$.

(b) Since $X \perp\!\!\!\perp Y \mid Z$, we have;

$$f(x, y, z)k(z) = g(x, z)p(y, z)$$

where $f(\cdot), k(\cdot), g(\cdot), p(\cdot)$ represents the respective density. Also, fix any u in the range of the random variable $h(X)$. Let us call, $S_u = \{x : h(x) = u\}$. Then, integrating x in both sides of the above equation over the set S_u, we obtain;

$$\begin{aligned}
&\int_{S_u} f(x, y, z)k(z)dx = \int_{S_u} g(x, z)p(y, z)dx \\
\Rightarrow\quad & k(z)\int_{S_u} f(h(X) = h, Y = y, Z = z) = p(y, z)\int_{S_u} g(x, z)dx \\
\Rightarrow\quad & f(h(X) = u, Y = y, Z = z)k(z) = p(y, z)g(h(X) = u, Z = z)
\end{aligned}$$

Hence, $U \perp\!\!\!\perp Y \mid Z$

(c) We shall use the generic symbol of $f(\cdot)$ to represent a density / p.m.f. Now, since $X \perp\!\!\!\perp Y \mid Z$, and $U = h(X)$, we can apply the result of part (b) to obtain $Y \perp\!\!\!\perp U \mid Z$. Hence, note that;

$$\begin{aligned}
f(y \mid z, u) &= \frac{f(y, u \mid z)}{f(u \mid z)} \\
&= \frac{f(y \mid z)f(u \mid z)}{f(u \mid z)} \quad \text{due to the independence from part (b)} \\
&= f(y \mid z)
\end{aligned}$$

Now, also note that, part (b) yields $(X, U) \perp\!\!\!\perp Y \mid Z$. Then consider the following;

$$\begin{aligned} f(x, y \mid z, u) &= \frac{f(x, y, u \mid z)}{f(u \mid z)} \\ &= \frac{f(x, u \mid z) f(y \mid z)}{f(u \mid z)} \\ &= \frac{f(x, u \mid z)}{f(u \mid z)} f(y \mid z, u) \quad \text{from the equality above} \\ &= f(x \mid z, u) f(y \mid z, u) \end{aligned}$$

Thus, $X \perp\!\!\!\perp Y \mid (Z, U)$. Note that, the conditional distributions exists since we have non-zero p.m.f. or density of conditioning variables, asserted by the given statements.

(d) We are given; $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp W \mid (Y, Z)$, which implies the following two factorizations;

$$\begin{aligned} f(x, y, z) k(z) &= g(x, z) h(y, z) \\ p(x, w, y, z) h(y, z) &= f(x, y, z) q(w, y, z) \end{aligned}$$

Now, note that,

$$\begin{aligned} & p(x, w, y, z) h(y, z) = f(x, y, z) q(w, y, z) \\ \Rightarrow \quad & p(x, w, y, z) h(y, z) k(z) = f(x, y, z) k(z) q(w, y, z) \\ \Rightarrow \quad & p(x, w, y, z) h(y, z) k(z) = g(x, z) h(y, z) q(w, y, z) \\ \Rightarrow \quad & p(x, w, y, z) k(z) = g(x, z) q(w, y, z) \end{aligned}$$

where the last line follows from the fact that $h(y, z) \neq 0$ for the particular choice of y and z, since it was conditioning variable in the given statement. The last line implies that $X \perp\!\!\!\perp (W, Y) \mid Z$.

(e) Since, $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp Z \mid Y$, we have;

$$f(x, y, z) k(z) = g(x, z) h(y, z) \Rightarrow f(x, y, z) = \frac{g(x, z) h(y, z)}{k(z)}$$

and

$$f(x, y, z) p(y) = q(x, y) h(y, z) \Rightarrow f(x, y, z) = \frac{q(x, y) h(y, z)}{p(y)}$$

which can be written because of positive density / p.m.f.

Combining the above, we write;

$$\begin{aligned} & \frac{g(x, z) h(y, z)}{k(z)} = \frac{q(x, y) h(y, z)}{p(y)} \\ \Rightarrow \quad & g(x, z) = \frac{q(x, y) k(z)}{p(y)} \quad \text{since } h(y, z) \neq 0 \end{aligned}$$

Since the left hand side does not depend on y, we can fix any value for y in the right hand side. Taking, $y = y_0$, we obtain;

$$g(x, z) = r(x, y_0) k(z)$$

where $r(x, y_0) = \frac{q(x, y_0)}{p(y_0)}$. Putting it back in the usual factorization, we get; $f(x, y, z) = \frac{g(x, z) h(y, z)}{k(z)} = \frac{r(x, y_0) k(z) h(y, z)}{k(z)} = r(x) k(z) h(y, z)$, therefore, $X \perp\!\!\!\perp (Y, Z)$. (Note that, since y_0 is just a constant, we may write $r(x, y_0) = r(x)$).

(f) Using a generic notation of $f(\cdot)$ for density, we write the given conditions;

$$f(x,y,w,z)f(w,z) = f(x,w,z)f(y,w,z)$$

and

$$f(x,y,w,z)f(y,z) = f(x,y,z)f(y,w,z)$$

Equating $f(x,y,w,z)$ from both the equation yields;

$$\frac{f(x,w,z)}{f(w,z)} = \frac{f(x,y,z)}{f(y,z)}$$

The left hand side does not depend on y, hence we can essentially fix $y = y_0$ at the right hand side as before. Therefore,

$$\frac{f(x,w,z)}{f(w,z)} = \frac{f(x,y_0,z)}{f(y_0,z)}$$

Now,

$$\begin{aligned} f(x,y,z,w)f(z) &= \frac{f(x,y,z,w)f(w,z)f(z)}{f(w,z)} \\ &= \frac{f(x,w,z)f(y,w,z)f(z)}{f(w,z)} \\ &= \frac{f(x,y_0,z)f(y,w,z)f(z)}{f(y_0,z)} \\ &= g(x,z)h(y,w,z) \end{aligned}$$

where $g(\cdot)$ and $h(\cdot)$ are suitably obtained by dropping y_0 from argument as it is a constant.

Thus, $X \perp\!\!\!\perp (W,Y) \mid Z$

To show that (e) and (f) are false without the assumption of positive p.m.f. or density, consider the following example for problem (e).

Take $X = Y = Z \sim Ber(0.5)$. Now, given Z, X and Y both are constants (basically degenerate random variable), and hence are independent, i.e. $X \perp\!\!\!\perp Y \mid Z$. By similar argument, $X \perp\!\!\!\perp Z \mid Y$. But, $X \not\perp\!\!\!\perp (Y,Z)$, since;

$$P(X=0,Y=1,Z=1) = 0 \neq P(X=0)P(Y=1,Z=1) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

To show a counterexample to problem (f), we take $X = Y = W \sim Ber(0.5)$, and $Z \sim Ber(0.5)$, independently of X, Y or W. Clearly, since Z is independent of others, the given situation completely reduces to problem (e), for which the above counterexample would work as shown.

2. Real data analysis with log-linear models $[(5+5) = 10]$

(i) Consider the `reinis` data from the **R** package `gRbase`. Perform backward and forward graphical model selection using AIC and BIC. Plot the graphs corresponding to the selected models. Identify the decomposable models among these and find an RIP ordering for each. Write down the conditional independences for each of the selected models. Compare the four selected models.

(ii) As in part (i), perform model selection on the data in Table 1. Interpret the selected models in terms of conditional independences and identify the decomposable ones among them. Implement the Iterative Proportional Scaling (IPS) procedure from scratch and using it compute the MLE of $m_{\text{Black, White, No}}$ (here the variables are ordered as "Victim's Race" (VR), "Defendant's Race" (DR), and "Death Penalty" (DP)) under each of the following hierarchical models:

Victim's Race	Defendant's Race	Death Penalty	
		Yes	No
White	White	19	132
	Black	11	52
Black	White	0	9
	Black	6	97

Table 1: Death penalty data

(a) $H = F_{\text{VR, DR}} + F_{\text{DR, DP}} + F_{\text{VR, DP}}$.

(b) $H = F_{\text{VR, DR}} + F_{\text{VR, DP}}$

Check that IPS converges in two iterations in case of model (b). How many iterations does it take to converge in case of model (a)? Now test the hypothesis H_0 : Model (b) vs H_1 : Model (a).

Solution. (i) We use the following code to load the library and the `reinis` data and take a look at it.

```
library(gRbase)
library(gRim)
data("reinis")
str(reinis)
```

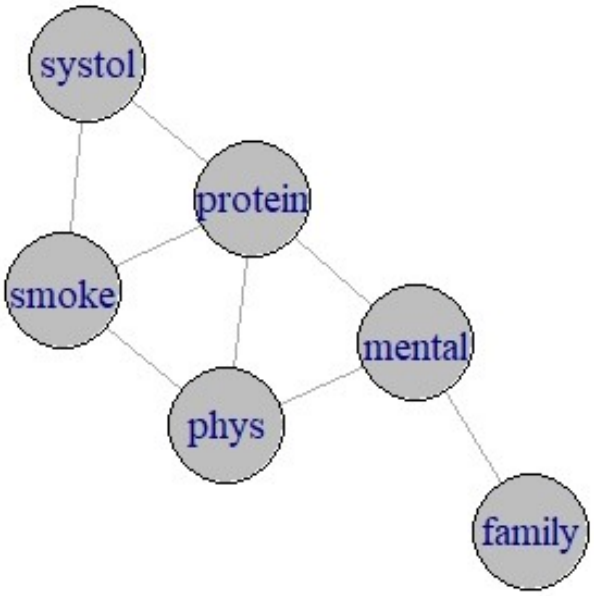
```
 'table' num [1:2, 1:2, 1:2, 1:2, 1:2, 1:2] 44 40 112 67 129 145 12 23 35 12 ...
 - attr(*, "dimnames")=List of 6
 ..$ smoke  : chr [1:2] "y" "n"
 ..$ mental : chr [1:2] "y" "n"
 ..$ phys   : chr [1:2] "y" "n"
 ..$ systol : chr [1:2] "y" "n"
 ..$ protein: chr [1:2] "y" "n"
 ..$ family : chr [1:2] "y" "n"
```

Now we fit a saturated model and an independence model in the data.

```
dm.sat <- dmod( ~ . ^ ., reinis) # fit saturated model
dm.null <- dmod( ~ . ^ 1, reinis)   # fit independence model
```

Firstly, we perform **AIC** based model selection.

```
# forward model selection using AIC starting from independence model,
   unrestricted to only decomposable models
fit.forward <- stepwise(dm.null, criterion = "aic", direction = "forward",
    type = "unrestricted", k = 2)
iplot(fit.forward)
```



Note that, the above chosen model is **decomposable**. The RIP ordering is as follows;

(a) (Systol, Smoke, Protein) as Clique 1.

(b) (Smoke, Protein, Phys) as Clique 2.

(c) (Protein, Phys, Mental) as Clique 3.

(d) (Mental, Family) as Clique 4.

The conditional independences are as follows;

(a) Systol $\perp\!\!\!\perp$ (Phys, Mental, Family) | (Protein, Smoke).

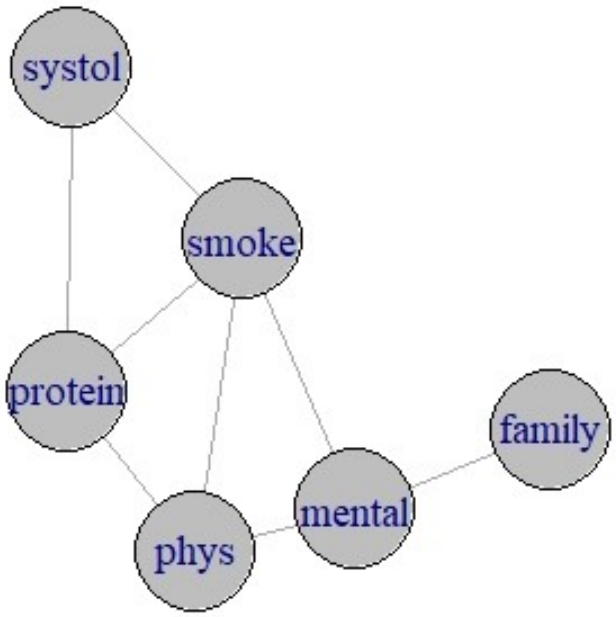
(b) (Systol, Smoke) $\perp\!\!\!\perp$ (Mental, Family) | (Protein, Phys).

(c) (Systol, Protein, Smoke, Phys) $\perp\!\!\!\perp$ Family | Mental.

All other conditional independences follow from taking subset of one or both of the independent sets and increasing the size of conditioning set. Hence, in a sense, all other conditional independences can be derived from the above relations.

The AIC of the above model is 13371.18 and BIC is 13464.99. Next, we perform backward model selection using AIC.

```
# backward model selection using AIC starting from saturated model
fit.backward <- stepwise(dm.sat, criterion = "aic", direction = "backward",
   type = "unrestricted", k = 2)
iplot(fit.backward)
```



Note that, this model is essentially very similar to the one before. It is decomposable. The RIP ordering is as follows;

(a) (Systol, Smoke, Protein) as Clique 1.

(b) (Smoke, Protein, Phys) as Clique 2.

(c) (Smoke, Phys, Mental) as Clique 3.

(d) (Mental, Family) as Clique 4.

The conditional independences are as follows;

(a) Systol $\perp\!\!\!\perp$ (Phys, Mental, Family) | (Protein, Smoke).

(b) (Systol, Protein) $\perp\!\!\!\perp$ (Mental, Family) | (Smoke, Phys).

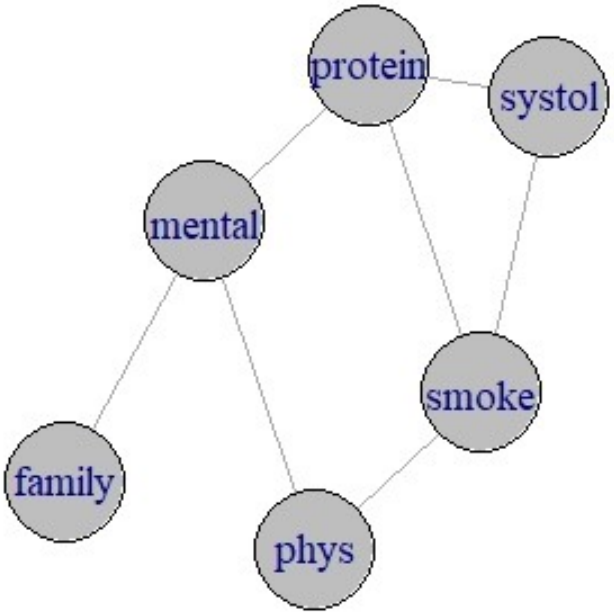
(c) (Systol, Protein, Smoke, Phys) $\perp\!\!\!\perp$ Family | Mental.

All other conditional independences follow from taking subset of one or both of the independent sets and increasing the size of conditioning set. Hence, in a sense, all other conditional independences can be derived from the above relations.

The AIC of the above model is 13371.63 and BIC is 13465.43, which is pretty close to the selected model before.

To perform BIC based model selection, we essentially use the same code, but with $k = \log n$, where n is the total number of observations. The **R** function `stepwise` uses the model selection criteria, $-2 \log L + kp$, where p is the number of parameters and k is the penalty parameter passed as an argument in the function.

```
1  # forward model selection using bic
2  fit.forward <- stepwise(dm.null, criterion = "aic", direction = "forward",
       type = "unrestricted", k = log(sum(reinis)))
3  iplot(fit.forward)
```



The above model is not decomposable, since (Protein, Smoke, Phys, Mental) is a chordless cycle of length 4. The AIC of the selected model is 13372.55 and the BIC is the selected model is 13449.8.

The conditional independences are as follows;

(a) Systol $\perp\!\!\!\perp$ (Phys, Mental, Family) | (Protein, Smoke).

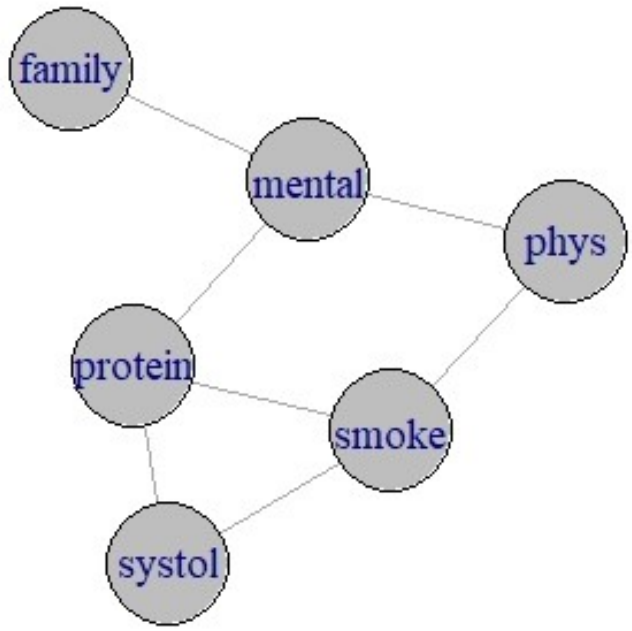
(b) (Systol, Smoke) $\perp\!\!\!\perp$ (Mental, Family) | (Protein, Phys).

(c) (Systol, Protein) $\perp\!\!\!\perp$ Family $\perp\!\!\!\perp$ Phys | (Smoke, Mental).

(d) (Systol, Protein, Smoke, Phys) $\perp\!\!\!\perp$ Family | Mental.

All other conditional independences follow from taking subset of one or both of the independent sets and increasing the size of conditioning set. Hence, in a sense, all other conditional independences can be derived from the above relations. Next, we use backward selection using BIC based criterion.

```
1    # backward model selection
2    fit.backward <- stepwise(dm.sat, criterion = "aic", direction = "backward"
        , type = "unrestricted", k = log(sum(reinis)))
3    iplot(fit.backward)
```



The selected model turns out to be same as the one selected by forward selection. Hence, it is not decomposable and the conditional independences are same as before.

In terms of AIC, the first model seems substantially better than the third model, while only marginally better than the second model. On the other hand, in terms of BIC, third model is substantially better than the first and second model. The first model being decomposable, is easy to work with for finding the maximum likelihood estimator.

(ii) Firstly, we load the dataset into **R** using the following code.

```
deathpenalty <- array(c(19, 132, 11, 52, 0, 9, 6, 97), dim = c(2,2,2),
   dimnames = list("Death.penalty" = c("Yes", "No"), "Defendant.race"= c("
   White","Black"), "Victim.race" = c("White","Black")))
ftable(deathpenalty)
```

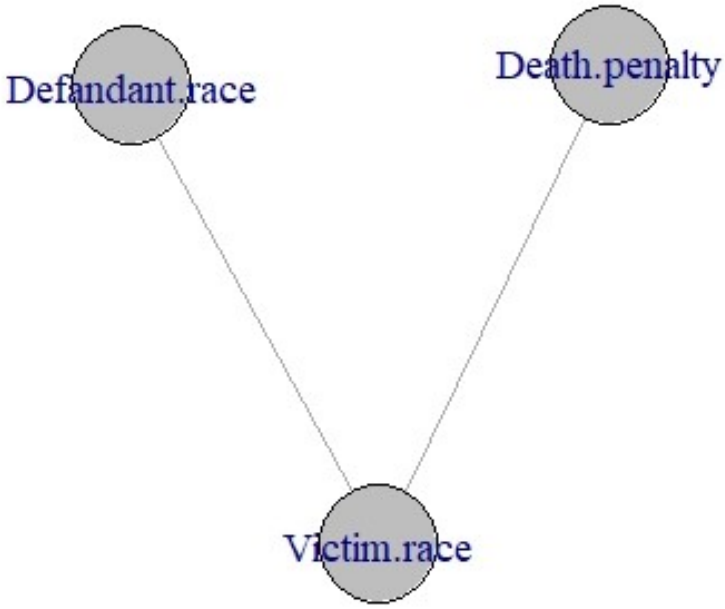
```
                          Victim.race White Black
Death.penalty Defendant.race
Yes           White                  19     0
              Black                  11     6
No            White                 132     9
              Black                  52    97
```

We first fit a saturated and an independence model for the data.

```
dm.sat <- dmod( ~ . ^ ., as.table(deathpenalty)) # fit saturated model
dm.null <- dmod( ~ . ^ 1, as.table(deathpenalty))   # fit independence
   model
```

We now perform AIC based forward and backward model selection.

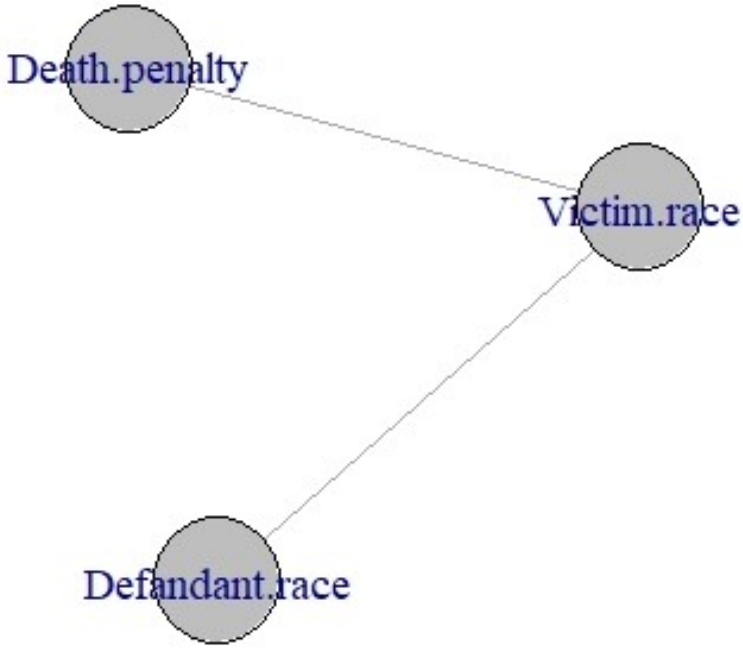
```
# forward model selection using AIC
fit.forward <- stepwise(dm.null, criterion = "aic", direction = "forward",
    type = "unrestricted", k = 2)
iplot(fit.forward)
```



The model is trivially decomposable. The AIC of the model is 971.7625 and BIC of the model is 990.697.

The only conditional independence in the model is that; Death Penalty ⫫ Defendant's Race | Victim's Race. Next, we apply the backward selection method.

```
# backward model selection using AIC
fit.backward <- stepwise(dm.sat, criterion = "aic", direction = "backward"
   , type = "unrestricted", k = 2)
iplot(fit.backward)
```



Note that, the same model is chosen again based on AIC criterion from backward selection.

To perform forward and backward model selection based on BIC criterion, we apply the same code again.

```
# forward model selection using BIC
fit.forward <- stepwise(dm.null, criterion = "aic", direction = "forward",
    type = "unrestricted", k = log(sum(deathpenalty)))
```

```
    iplot(fit.forward)

    # backward model selection using BIC
    fit.backward <- stepwise(dm.sat, criterion = "aic", direction = "backward"
       , type = "unrestricted", k = log(sum(deathpenalty)))
    iplot(fit.backward)
```

Both the above selected model gives the same model as above.

Now, following is the code for Iterative Proportional Scaling (IPS) Algorithm implemented from scratch using **R**. We use the package **gRbase** for different utility functions such as slicing and marginalization.

```
    IPS <- function(form, data, maxiter = 1000, tol = 1e-05) {
    # initialize an array of 1's only
    tempdata <- array(1, dim = dim(data), dimnames = dimnames(data))
    form.vars <- labels(terms(formula(form)))

    current_error <- Inf
    current_iter <- 0
    while (current_error > tol & current_iter < maxiter) {
    # store the current table for computing the error later
    current_tab <- tempdata

    # for each given margin, perform the updation
    for (var in rev(form.vars))   {
    true_margins <- ar_marg(data, formula(paste0("~", var)))   # compute the
       true marginals
    current_margins <- ar_marg(tempdata, formula(paste0("~", var)))   #
       compute the current margins

    # expand them to higher dimension as original so that multiplication can
       be performed
    true_margins <- ar_expand(true_margins, dimnames(data))
    current_margins <- ar_expand(current_margins, dimnames(data))

    tempdata <- (tempdata * true_margins)/current_margins
    }

    current_iter <- current_iter+ 1        # increase number of iteration
    current_error <- max(abs(tempdata - current_tab))   # compute the error
    }

    return(list("MLE Table" = tempdata, "Iteration" = current_iter))
    }
```

Now, we call the above function to find maximum likelihood estimators under model (a).

```
    IPS( ~ Victim.race + Death.penalty + Defendant.race + Victim.race:Death.
       penalty +
    Death.penalty:Defendant.race + Victim.race:Defendant.race, deathpenalty)
```

```
$`MLE Table`
, , Victim.race = White

Defendant.race
Death.penalty       White       Black
Yes                 18.67436    11.32564
No                  132.32564   51.67436

, , Victim.race = Black

Defendant.race
Death.penalty       White       Black
Yes                 0.3256423   5.674358
No                  8.6743584   97.325642

$Iteration
[1] 11
```

Note that, the maximum likelihood table is extremely close to the true table. The maximum likelihood estimate of $m_{\text{Black, White, No}} = 8.674$, which is pretty close to the true value.

Now, we call the above function again to find maximum likelihood estimators under model (b).

```
IPS( ~ Victim.race + Death.penalty + Defendant.race + Victim.race:Defendant.
   race + Victim.race:Death.penalty, deathpenalty)
```

```
$`MLE Table`
, , Victim.race = White

Defendant.race
Death.penalty       White       Black
Yes                 21.16822    8.831776
No                  129.83178   54.168224

, , Victim.race = Black

Defendant.race
Death.penalty       White       Black
Yes                 0.4821429   5.517857
No                  8.5178571   97.482143

$Iteration
[1] 2
```

The maximum likelihood estimate of $m_{\text{Black, White, No}} = 8.5178$, which is pretty close to the true value of 9.

Note that, IPS converges in two iterations for the case of model (b).

Now, to test the hypothesis H_0 : Model (b) vs H_1 : Model (a), we use a likelihood ratio test.

For H_0, under model (b), we have the value of log likelihood is (-480.8812) and the corresponding number of free parameters is 5. While, under H_1, model (a), the value of log likelihood turns out to be (-480.2907), which 6 free parameters. Hence,

$$-2\log\lambda = -2\log\frac{L_{H_0}}{L_{H_0\cup H_1}} = -2(\log L_{H_0} - \log L_{H_1}) = 1.181$$

since, $H_0 \cup H_1 = H_1$, as $H_0 \subset H_1$. Note that, the above likelihood ratio statistic asymptotically follows central chi-squared distribution with degrees of freedom $(6-5) = 1$. The observed chi-square statistic assumes a p-value of 0.2771525, which is greater than the significance level of $\alpha = 0.05$, hence we fail to reject the null hypothesis in favour of the alternative and richer model (a).

Hence, we can accept Model (b) as an possibly appropriate model for the given data. Also note that, this model is suggested based on AIC or BIC criterion above.