# Nonparametric Estimation of IV Model

Jiaqi Yin

May 22, 2018

## 1 Estimations

We consider binary instrumental variable model Fig.(1) , where we have $A \perp U$, $C \perp A \mid (B, U)$ and $A$ is not independent of $B$. We have $A, B, C \in \{0, 1\}$. Let $P_{cb \cdot a} = Pr(C = c, B = b \mid A = a)$. Sufficient and necessary conditions for IV model are

$$\begin{cases} P_{00 \cdot 0} + & P_{10 \cdot 1} \leq 1 \\ P_{01 \cdot 0} + & P_{11 \cdot 1} \leq 1 \\ P_{10 \cdot 0} + & P_{00 \cdot 1} \leq 1 \\ P_{11 \cdot 0} + & P_{01 \cdot 1} \leq 1 \end{cases} . \tag{1}$$

Ineq.(1) are also called IV-inequalites.

Suppose our sample size is $n$, and each individual is i.i.d. Let $n_{cba} = \sum_{i=1}^{n} \mathbb{I}\{A = a, B = b, C = c\}$ . We need to esitimate $P_{cb \cdot a}$. The log-likelihood is

$$l(P_{cb \cdot a}) = \sum_{a,b,c} n_{cba} \log P_{cb \cdot a}.$$

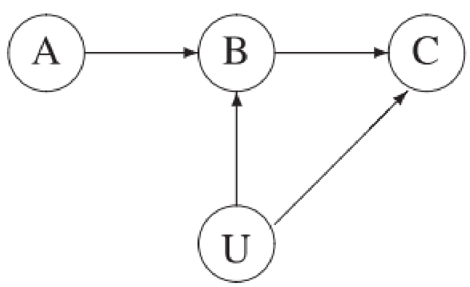


Figure 1: Directed acyclic graph which represents the instrumental variable model.

Of cousse, we assume $P_{cb\cdot a} > 0$. An optimization question is raised as follows,

$$\begin{cases} \min & -\sum_{a,b,c} n_{cba} \log P_{cb\cdot a} \\ \text{subject to} & P_{00\cdot 0} + P_{10\cdot 1} \leq 1 \\ & P_{01\cdot 0} + P_{11\cdot 1} \leq 1 \\ & P_{10\cdot 0} + P_{00\cdot 1} \leq 1 \\ & P_{11\cdot 0} + P_{01\cdot 1} \leq 1 \\ & P_{00\cdot 0} + P_{01\cdot 0} + P_{10\cdot 0} + P_{11\cdot 0} = 1 \\ & P_{00\cdot 1} + P_{01\cdot 1} + P_{10\cdot 1} + P_{11\cdot 1} = 1 \end{cases},$$

where objective function and feasible set are convex. The Lagrange fucntion is

$$\begin{aligned} L(P_{abc}, \lambda, \nu) = &-\sum_{a,b,c} n_{cba} \log P_{cb\cdot a} + \lambda_1 \left(P_{00\cdot 0} + P_{10\cdot 1} - 1\right) + \lambda_2 \left(P_{01\cdot 0} + P_{11\cdot 1} - 1\right) \\ &+ \lambda_3 \left(P_{10\cdot 0} + P_{00\cdot 1} - 1\right) + \lambda_4 \left(P_{11\cdot 0} + P_{01\cdot 1} - 1\right) \\ &+ \nu_1 \left(P_{00\cdot 0} + P_{01\cdot 0} + P_{10\cdot 0} + P_{11\cdot 0} - 1\right) + \nu_2 \left(P_{00\cdot 1} + P_{01\cdot 1} + P_{10\cdot 1} + P_{11\cdot 1} - 1\right). \end{aligned}$$

Becasue Slater's condition is satisfied, strong duality holds. Let $p^*_{cb\cdot a}$ and $(\lambda^*, \nu^*)$ be primal and dual optimal points. We further have KKT conditions,

$$\begin{cases} p^*_{00\cdot 0} + p^*_{10\cdot 1} - 1 \leq 0 \\ p^*_{01\cdot 0} + p^*_{11\cdot 1} - 1 \leq 0 \\ p^*_{10\cdot 0} + p^*_{00\cdot 1} - 1 \leq 0 \\ p^*_{11\cdot 0} + p^*_{01\cdot 1} - 1 \leq 0 \\ p^*_{00\cdot 0} + p^*_{01\cdot 0} + p^*_{10\cdot 0} + p^*_{11\cdot 0} - 1 = 0 \\ p^*_{00\cdot 1} + p^*_{01\cdot 1} + p^*_{10\cdot 1} + p^*_{11\cdot 1} - 1 = 0 \\ \lambda^*_i \geq 0, \quad i = 1,2,3,4 \\ \lambda^*_1 \left(p^*_{00\cdot 0} + p^*_{10\cdot 1} - 1\right) = 0 \\ \lambda^*_2 \left(p^*_{01\cdot 0} + p^*_{11\cdot 1} - 1\right) = 0 \\ \lambda^*_3 \left(p^*_{10\cdot 0} + p^*_{00\cdot 1} - 1\right) = 0 \\ \lambda^*_4 \left(p^*_{11\cdot 0} + p^*_{01\cdot 1} - 1\right) = 0 \\ -\frac{n_{000}}{p^*_{00\cdot 0}} + \lambda^*_1 + \nu^*_1 = 0; \quad -\frac{n_{101}}{p^*_{10\cdot 1}} + \lambda^*_1 + \nu^*_2 = 0; \\ -\frac{n_{010}}{p^*_{01\cdot 0}} + \lambda^*_2 + \nu^*_1 = 0; \quad -\frac{n_{111}}{p^*_{11\cdot 1}} + \lambda^*_2 + \nu^*_2 = 0; \\ -\frac{n_{100}}{p^*_{10\cdot 0}} + \lambda^*_3 + \nu^*_1 = 0; \quad -\frac{n_{001}}{p^*_{00\cdot 1}} + \lambda^*_3 + \nu^*_2 = 0; \\ -\frac{n_{110}}{p^*_{11\cdot 0}} + \lambda^*_4 + \nu^*_1 = 0; \quad -\frac{n_{011}}{p^*_{01\cdot 1}} + \lambda^*_4 + \nu^*_2 = 0. \end{cases}$$

We start by noting that $\lambda^*$act as slack variables in the last four equations, so it can be eliminated, leaving

$$
\begin{cases}
p^*_{00\cdot 0} + p^*_{10\cdot 1} - 1 \leq 0 \\
p^*_{01\cdot 0} + p^*_{11\cdot 1} - 1 \leq 0 \\
p^*_{10\cdot 0} + p^*_{00\cdot 1} - 1 \leq 0 \\
p^*_{11\cdot 0} + p^*_{01\cdot 1} - 1 \leq 0 \\
p^*_{00\cdot 0} + p^*_{01\cdot 0} + p^*_{10\cdot 0} + p^*_{11\cdot 0} - 1 = 0 \\
p^*_{00\cdot 1} + p^*_{01\cdot 1} + p^*_{10\cdot 1} + p^*_{11\cdot 1} - 1 = 0 \\
\left(\frac{n_{000}}{p^*_{00\cdot 0}} - \nu^*_1\right)(p^*_{00\cdot 0} + p^*_{10\cdot 1} - 1) = 0; \quad \left(\frac{n_{101}}{p^*_{10\cdot 1}} - \nu^*_2\right)(p^*_{00\cdot 0} + p^*_{10\cdot 1} - 1) = 0 \\
\left(\frac{n_{010}}{p^*_{01\cdot 0}} - \nu^*_1\right)(p^*_{01\cdot 0} + p^*_{11\cdot 1} - 1) = 0; \quad \left(\frac{n_{111}}{p^*_{11\cdot 1}} - \nu^*_2\right)(p^*_{01\cdot 0} + p^*_{11\cdot 1} - 1) = 0 \\
\left(\frac{n_{100}}{p^*_{10\cdot 0}} - \nu^*_1\right)(p^*_{10\cdot 0} + p^*_{00\cdot 1} - 1) = 0; \quad \left(\frac{n_{001}}{p^*_{00\cdot 1}} - \nu^*_2\right)(p^*_{10\cdot 0} + p^*_{00\cdot 1} - 1) = 0 \\
\left(\frac{n_{110}}{p^*_{11\cdot 0}} - \nu^*_1\right)(p^*_{11\cdot 0} + p^*_{01\cdot 1} - 1) = 0; \quad \left(\frac{n_{011}}{p^*_{01\cdot 1}} - \nu^*_2\right)(p^*_{11\cdot 0} + p^*_{01\cdot 1} - 1) = 0 \\
\nu^*_1 \leq \min\left\{\frac{n_{000}}{p^*_{00\cdot 0}}, \frac{n_{010}}{p^*_{01\cdot 0}}, \frac{n_{100}}{p^*_{10\cdot 0}}, \frac{n_{110}}{p^*_{11\cdot 0}}\right\} \\
\nu^*_2 \leq \min\left\{\frac{n_{101}}{p^*_{10\cdot 1}}, \frac{n_{111}}{p^*_{11\cdot 1}}, \frac{n_{001}}{p^*_{00\cdot 1}}, \frac{n_{011}}{p^*_{01\cdot 1}}\right\}
\end{cases}.
$$

If $p^*_{00\cdot 0} + p^*_{10\cdot 1} - 1 < 0; p^*_{01\cdot 0} + p^*_{11\cdot 1} - 1 < 0; p^*_{10\cdot 0} + p^*_{00\cdot 1} - 1 < 0; p^*_{11\cdot 0} + p^*_{01\cdot 1} - 1 < 0$, we will have

$$
\begin{aligned}
\nu^*_1 = \frac{n_{000}}{p^*_{00\cdot 0}} = \frac{n_{010}}{p^*_{01\cdot 0}} = \frac{n_{100}}{p^*_{10\cdot 0}} = \frac{n_{110}}{p^*_{11\cdot 0}}, \\
\nu^*_2 = \frac{n_{101}}{p^*_{10\cdot 1}} = \frac{n_{111}}{p^*_{11\cdot 1}} = \frac{n_{001}}{p^*_{00\cdot 1}} = \frac{n_{011}}{p^*_{01\cdot 1}}.
\end{aligned}
$$

Further, $\nu^*_1 = n_0$, $\nu^*_2 = n_1$, and $p^*_{cb\cdot a} = \frac{n_{cba}}{n_a}$.
However, once we consider sampling variability, we could have the case such that $\frac{n_{cba}}{n_a} + \frac{n_{(1-c)b(1-a)}}{n_{1-a}} > 1$, which voilates IV Inqualities if we estimate $p^*_{cb\cdot a} = \frac{n_{cba}}{n_a}$. When we have such case, let $p^*_{c'b'\cdot a'} + p^*_{(1-c')b'\cdot(1-a')} = 1$ for some $a', b', c'$. Consider nonzero $p^*_{cb\cdot a}$ and $\sum_{bc} p^*_{cb\cdot a} = 1$, we could not have $p^*_{(1-c')b'\cdot a'} + p^*_{c'b'\cdot(1-a')} = 1$, $p^*_{(1-c')(1-b')\cdot a'} + p^*_{c'(1-b')\cdot(1-a')} = 1$ or $p^*_{c'(1-b')\cdot a'} + p^*_{(1-c')(1-b')\cdot(1-a')} = 1$ (once any one of it holds, we will have zero proabability). WOLG, let $c' = 0, b' = 0, a' = 0$, and $\frac{n_{000}}{n_0} + \frac{n_{101}}{n_1} > 1$. Let $p^*_{00\cdot 0} + p^*_{10\cdot 1} = 1$, $p^*_{01\cdot 0} + p^*_{11\cdot 1} < 1$, $p^*_{10\cdot 0} + p^*_{00\cdot 1} < 1$, and $p^*_{11\cdot 0} + p^*_{01\cdot 1} < 1$.
Further,

$$
\begin{aligned}
\frac{n_{010}}{p^*_{01\cdot 0}} = \frac{n_{100}}{p^*_{10\cdot 0}} = \frac{n_{110}}{p^*_{11\cdot 0}} = \nu^*_1 < \frac{n_{000}}{p^*_{00\cdot 0}} \\
\frac{n_{111}}{p^*_{11\cdot 1}} = \frac{n_{001}}{p^*_{00\cdot 1}} = \frac{n_{011}}{p^*_{01\cdot 1}} = \nu^*_2 < \frac{n_{101}}{p^*_{10\cdot 1}}.
\end{aligned} \tag{2}
$$

The constraints are updated as

$$
\begin{cases}
p^*_{00\cdot 0} + p^*_{10\cdot 1} = 1 \\
p^*_{01\cdot 0} + p^*_{11\cdot 1} + p^*_{10\cdot 0} + p^*_{00\cdot 1} + p^*_{11\cdot 0} + p^*_{01\cdot 1} = 1 \\
\frac{n_{010}}{p^*_{01\cdot 0}} = \frac{n_{100}}{p^*_{10\cdot 0}} = \frac{n_{110}}{p^*_{11\cdot 0}} = \nu^*_1 < \frac{n_{000}}{p^*_{00\cdot 0}} \\
\frac{n_{111}}{p^*_{11\cdot 1}} = \frac{n_{001}}{p^*_{00\cdot 1}} = \frac{n_{011}}{p^*_{01\cdot 1}} = \nu^*_2 < \frac{n_{101}}{p^*_{10\cdot 1}} \\
p^*_{01\cdot 0} + p^*_{10\cdot 0} + p^*_{11\cdot 0} < 1 \\
p^*_{01\cdot 1} + p^*_{10\cdot 1} + p^*_{11\cdot 1} < 1
\end{cases} . \tag{3}
$$

From Eq.(3)-1,2, we have $p^*_{00\cdot 0} = 1 - p^*_{10\cdot 1} = p^*_{00\cdot 1} + p^*_{11\cdot 0} + p^*_{01\cdot 1}$, and $p^*_{10\cdot 1} = 1 - p^*_{00\cdot 0} = p^*_{01\cdot 0} + p^*_{11\cdot 1} + p^*_{10\cdot 0}$. Plug them into (3)-3,4, we further have

$$
\frac{n_{010}}{p^*_{01\cdot 0}} = \frac{n_{100}}{p^*_{10\cdot 0}} = \frac{n_{110}}{p^*_{11\cdot 0}} = \frac{n_0 - n_{000}}{1 - p^*_{00\cdot 0}} = \frac{n_0 - n_{000}}{p^*_{10\cdot 1}}. \tag{4}
$$

$$
\frac{n_{111}}{p^*_{11\cdot 1}} = \frac{n_{001}}{p^*_{00\cdot 1}} = \frac{n_{011}}{p^*_{01\cdot 1}} = \frac{n_1 - n_{101}}{1 - p^*_{10\cdot 1}} = \frac{n_1 - n_{101}}{p^*_{00\cdot 0}}. \tag{5}
$$

From above equations, it yelds

$$
p^*_{01\cdot 0} = \frac{n_{010}}{n_0 - n_{000}} p^*_{10\cdot 1}, \quad p^*_{10\cdot 0} = \frac{n_{100}}{n_0 - n_{000}} p^*_{10\cdot 1}, \quad p^*_{11\cdot 0} = \frac{n_{110}}{n_0 - n_{000}} p^*_{10\cdot 1}; \tag{6}
$$

$$
p^*_{11\cdot 1} = \frac{n_{111}}{n_1 - n_{101}} p^*_{00\cdot 0}, \quad p^*_{00\cdot 1} = \frac{n_{001}}{n_1 - n_{101}} p^*_{00\cdot 0}, \quad p^*_{01\cdot 1} = \frac{n_{011}}{n_1 - n_{101}} p^*_{00\cdot 0}. \tag{7}
$$

From Ineq.(3)-3,4,5,6, we further have

$$
\begin{cases}
\frac{n_0 - n_{000}}{1 - p^*_{00\cdot 0}} < \frac{n_{000}}{p^*_{00\cdot 0}} \\
\frac{n_1 - n_{101}}{1 - p^*_{10\cdot 1}} < \frac{n_{101}}{p^*_{10\cdot 1}} \\
\left( \frac{n_{010}}{n_0 - n_{000}} + \frac{n_{100}}{n_0 - n_{000}} + \frac{n_{110}}{n_0 - n_{000}} \right) p^*_{10\cdot 1} < 1 \\
\left( \frac{n_{111}}{n_1 - n_{101}} + \frac{n_{001}}{n_1 - n_{101}} + \frac{n_{011}}{n_1 - n_{101}} \right) p^*_{00\cdot 0} < 1
\end{cases} .
$$

Finally, we have

$$
\max \left\{ 0, 1 - \frac{n_{101}}{n_1} \right\} < p^*_{00\cdot 0} < \min \left\{ \frac{n_{000}}{n_0}, 1 \right\} \tag{8}
$$

Next, we need to maximize $l(p^*_{cb\cdot a}) = \sum_{a,b,c} n_{cba} \log p^*_{cb\cdot a}$, and we further plug Eq.(6)(7) into the log-likelihood

| A | B | C | Count |
|---|---|---|---|
| 0 | 0 | 0 | 150 |
| 0 | 1 | 0 | 50 |
| 0 | 0 | 1 | 100 |
| 0 | 1 | 1 | 200 |
| 1 | 0 | 0 | 50 |
| 1 | 1 | 0 | 325 |
| 1 | 0 | 1 | 100 |
| 1 | 1 | 1 | 25 |

Table 1: fake data set

$$
\begin{aligned}
l(p^*_{00\cdot 0}) &= \sum_{a,b,c} n_{cba} \log p^*_{cb\cdot a} \\
&= n_{101} \log p^*_{10\cdot 1} + n_{010} \log\left(\frac{n_{010}}{n_0 - n_{000}} p^*_{10\cdot 1}\right) + n_{100} \log\left(\frac{n_{100}}{n_0 - n_{000}} p^*_{10\cdot 1}\right) + n_{110} \log\left(\frac{n_{110}}{n_0 - n_{000}} p^*_{10\cdot 1}\right) \\
&\quad + n_{000} \log p^*_{00\cdot 0} + n_{111} \log\left(\frac{n_{111}}{n_1 - n_{101}} p^*_{00\cdot 0}\right) + n_{001} \log\left(\frac{n_{001}}{n_1 - n_{101}} p^*_{00\cdot 0}\right) + n_{011} \log\left(\frac{n_{011}}{n_1 - n_{101}} p^*_{00\cdot 0}\right) \\
&= (n_{101} + n_{010} + n_{100} + n_{110}) \log(1 - p^*_{00\cdot 0}) + n_{010} \log\left(\frac{n_{010}}{n_0 - n_{000}}\right) + n_{100} \log\left(\frac{n_{100}}{n_0 - n_{000}}\right) + n_{110} \log\left(\frac{n_{110}}{n_0 - n_{000}}\right) \\
&\quad + (n_{000} + n_{111} + n_{001} + n_{011}) \log p^*_{00\cdot 0} + n_{111} \log\left(\frac{n_{111}}{n_1 - n_{101}}\right) + n_{001} \log\left(\frac{n_{001}}{n_1 - n_{101}}\right) + n_{011} \log\left(\frac{n_{011}}{n_1 - n_{101}}\right)
\end{aligned}
$$

The derivative of $l(p^*_{00\cdot 0})$ is

$$
l^{'}(p^*_{00\cdot 0}) = \frac{n_{000} + n_{111} + n_{001} + n_{011}}{p^*_{00\cdot 0}} - \frac{n_{101} + n_{010} + n_{100} + n_{110}}{1 - p^*_{10\cdot 0}}.
$$

Let $l^{'}(p^*_{00\cdot 0}) = 0$, we have maximizer $\tilde{p}_{00\cdot 0} = \frac{n_{000}+n_{111}+n_{001}+n_{011}}{n}$. When $p^*_{00\cdot 0} < \tilde{p}_{00\cdot 0}$, $l(p^*_{00\cdot 0})$ increases; when $p^*_{00\cdot 0} > \tilde{p}_{00\cdot 0}$, $l(p^*_{00\cdot 0})$ decreases.
Denote the set of $p^*_{00\cdot 0}$ in (8) as $\mathcal{C}$. Follow the following steps to have $p^*_{00\cdot 0}$. First, according to the data, have the constraint set $\mathcal{C}$ of $p^*_{00\cdot 0}$; second, $p^*_{00\cdot 0} = \arg\min_{p\in\mathcal{C}} |\tilde{p}_{00\cdot 0} - p|$.

## 2 Simulation

Suppose we have dataset (fake).
First, we calculate the empirical distribution

$$
\begin{aligned}
\pi_{emp} &= (\hat{p}_{00,0}, \hat{p}_{10,0}, \hat{p}_{01,0}, \hat{p}_{11,0}, \hat{p}_{00,1}, \hat{p}_{10,1}, \hat{p}_{01,1}, \hat{p}_{11,1}) \\
&= \left(\frac{150}{500}, \frac{50}{500}, \frac{100}{500}, \frac{200}{500}, \frac{50}{500}, \frac{325}{500}, \frac{100}{500}, \frac{25}{500}\right) \\
&= (0.3, 0.1, 0.2, 0.4, 0.1, 0.65, 0.2, 0.05).
\end{aligned}
$$

Find that $\hat{p}_{11,0} + \hat{p}_{01,1} = 0.4 + 0.65 = 1.05 > 1$, which voilates the IV-inequality. After applying Section 1 result, we have the restricted estimation of probability distribution as

$$
\begin{aligned}
\tilde{\pi}_{obs} &= \left(\tilde{p}_{00,0}, \tilde{p}_{10,0}, \tilde{p}_{01,0}, \tilde{p}_{11,0}, \tilde{p}_{00,1}, \tilde{p}_{10,1}, \tilde{p}_{01,1}, \tilde{p}_{11,1}\right) \\
&= (0.312, 0.104, 0.208, 0.375, 0.107, 0.625, 0.214, 0.054).
\end{aligned}
$$

We further generate data according to multinomial distribution $(\tilde{n}_{000}, \tilde{n}_{100}, \tilde{n}_{010}, \tilde{n}_{110}) \sim Multinom\,(n_0, (\tilde{p}_{00,0}, \tilde{p}_{10,0}, \tilde{p}_{01,0}, \tilde{p}_{11,0}))$, $(\tilde{n}_{001}, \tilde{n}_{101}, \tilde{n}_{011}, \tilde{n}_{111}) \sim Multinom\,(n_1, (\tilde{p}_{00,1}, \tilde{p}_{10,1}, \tilde{p}_{01,1}, \tilde{p}_{11,1}))$. The null hypothesis

$$H_0 : \text{the observed data from IV model.}$$

The simulation size is 10,000. We define the statistics as

$$\Lambda = 2\left[\ln(\text{likelihood for alternative model) -}\ln\left(\text{likelihood for null model}\right)\right].$$

The null model here is the restricted model, and alternative model is the emipirical model. We denote $\Lambda\left(\pi', \tilde{\pi}_{obs}\right) = 2\left(\sum_{a,b,c} n'_{cba} \log \frac{n'_{cba}}{n_a} - \sum_{a,b,c} n_{cba} \log \tilde{p}_{cb\cdot a}\right)$ in order to emphasize the parameters (distributions). We finally have

$$P\left(\Lambda\left(\pi', \tilde{\pi}_{obs}\right) \geq \Lambda(\pi_{emp}, \tilde{\pi}_{obs})\right) = 0.1168$$

# 3 Average Causal Effect

$$
\text{ACE}\,(D \to Y) \geq \max \left\{ \begin{array}{c}
p_{00\cdot 0} + p_{11\cdot 1} - 1 \\
p_{00\cdot 1} + p_{11\cdot 1} - 1 \\
p_{11\cdot 0} + p_{00\cdot 1} - 1 \\
p_{00\cdot 0} + p_{11\cdot 0} - 1 \\
2p_{00\cdot 0} + p_{11\cdot 0} + p_{10\cdot 1} + p_{11\cdot 1} - 2 \\
p_{00\cdot 0} + 2p_{11\cdot 0} + p_{00\cdot 1} + p_{01\cdot 1} - 2 \\
p_{10\cdot 0} + p_{11\cdot 0} + 2p_{00\cdot 1} + p_{11\cdot 1} - 2 \\
p_{00\cdot 0} + p_{01\cdot 0} + p_{00\cdot 1} + 2p_{11\cdot 1} - 2
\end{array} \right\}
$$

$$
\text{ACE}\,(D \to Y) \leq \min \left\{ \begin{array}{c}
1 - p_{10\cdot 0} + p_{01\cdot 1} \\
1 - p_{01\cdot 0} + p_{10\cdot 1} \\
1 - p_{01\cdot 0} + p_{10\cdot 0} \\
1 - p_{01\cdot 1} + p_{10\cdot 1} \\
2 - 2p_{01\cdot 0} - p_{10\cdot 0} - p_{10\cdot 1} - p_{11\cdot 1} \\
2 - p_{01\cdot 0} - 2p_{10\cdot 0} - p_{00\cdot 1} - p_{01\cdot 1} \\
2 - p_{10\cdot 0} - p_{11\cdot 0} - 2p_{01\cdot 1} - p_{10\cdot 1} \\
2 - p_{00\cdot 0} - p_{01\cdot 0} - p_{01\cdot 1} - 2p_{10\cdot 1}
\end{array} \right\}
$$