Nondifferential Exposure Misclassification in Case-Control Studies: What should be done with the 'Maybe' Exposed?

Dongxu Wang

Department of Statistics
UBC

2012/03

Outline

Regular 2×2 Table

Interested Cases: Unmatched case-control study.

		Apparent Exposure Status	
		Absence (0)	Presence (1)
Disease Status	Absence (0)	n_1	n_2
	Presence (1)	n_3	n_4

Exposure Misclassification

Let:
Y be the disease status;
X be the actual exposure status;
X^* be the apparent exposure status.

- bias estimates of exposure-disease association, i.e. misleading relative risk and OR

Two types of misclassification:

- **nondifferential misclassification**
- differential misclassification

2×3 Table

When the exposure prevalence is low, if the Exposure Status is not Sure, classify it as unexposed.
New Idea: thinking of X^* has three categories, unlikely exposed (0), maybe exposed (1), and likely exposed (2).

		X^*		
		0	1	2
Y	0	n_1	n_2	n_3
	1	n_4	n_5	n_6

2×3 Table

The prevalences of true exposure among controls and cases (goal of the study):

$$r_0 = Pr\{X = 1 \mid Y = 0\},$$

$$r_1 = Pr\{X = 1 \mid Y = 1\}.$$

The probability of misclassification:

$$\boldsymbol{p_0} = \begin{pmatrix} p_{00} \\ p_{01} \\ p_{02} \end{pmatrix} = \begin{pmatrix} Pr\{X^* = 0 \mid X = 0\} \\ Pr\{X^* = 1 \mid X = 0\} \\ Pr\{X^* = 2 \mid X = 0\} \end{pmatrix},$$

$$\boldsymbol{p_1} = \begin{pmatrix} p_{10} \\ p_{11} \\ p_{12} \end{pmatrix} = \begin{pmatrix} Pr\{X^* = 0 \mid X = 1\} \\ Pr\{X^* = 1 \mid X = 1\} \\ Pr\{X^* = 2 \mid X = 1\} \end{pmatrix}.$$

2×3 Table

The prevalences of apparent exposure among controls and cases (observed from data):

$$\boldsymbol{\theta_0} = r_0 \boldsymbol{p_1} + (1 - r_0)\boldsymbol{p_0} = \begin{pmatrix} Pr\{X^* = 0 \mid Y = 0\} \\ Pr\{X^* = 1 \mid Y = 0\} \\ Pr\{X^* = 2 \mid Y = 0\} \end{pmatrix},$$

$$\boldsymbol{\theta_1} = r_1 \boldsymbol{p_1} + (1 - r_1)\boldsymbol{p_0} = \begin{pmatrix} Pr\{X^* = 0 \mid Y = 1\} \\ Pr\{X^* = 1 \mid Y = 1\} \\ Pr\{X^* = 2 \mid Y = 1\} \end{pmatrix}.$$

Three Categories for Exposure

To ensure the classification is better than random, consider two assumptions:

- Constraint A: $p_{02} < p_{12}$ and $p_{00} > p_{10}$.
- Constraint B: $p_{00} > p_{01} > p_{02}$ and $p_{10} < p_{11} < p_{12}$.

Without loss of generality, assume that $r_0 < r_1$.

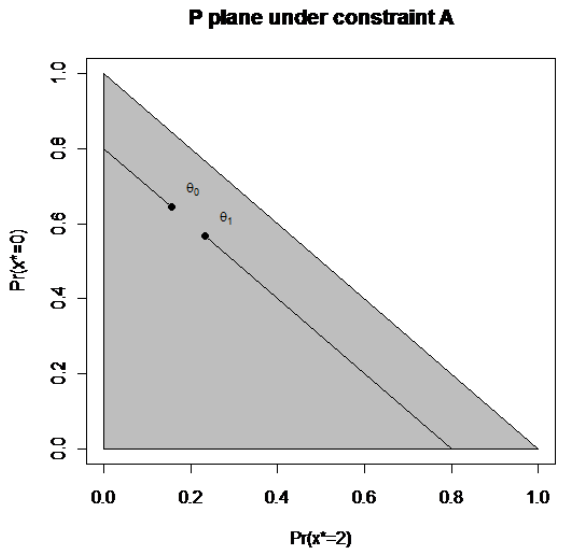
Prior Region and Identification Region

- Prior region $\mathbb{P}$: $\boldsymbol{p} = (\boldsymbol{p_0}, \boldsymbol{p_1})$
- Identification region $\mathbb{Q}(\theta)$: Given θ, all values of the target parameters (r_0, r_1) which yield this value of θ for some choice of $\boldsymbol{p} \in \mathbb{P}$.

Constraint A

The identification region in the $\boldsymbol{p}$ plane:

$$\begin{aligned}\{\boldsymbol{p_0}, \boldsymbol{p_1} :& \theta_{00} < p_{00} < 1, 0 < p_{02} < \theta_{02}, \\ & 0 < p_{10} < \theta_{10}, \theta_{12} < p_{12} < 1, \\ & p_{00} = a p_{02} + b, p_{10} = a p_{12} + b\},\end{aligned}$$



Constraint A

To transform the identification region in the $\boldsymbol{p}$ plane to the $\boldsymbol{r}$ plane, setting $z_0 = r_0/(r_1 - r_0)$ and $z_1 = (1 - r_1)/(r_1 - r_0)$.

$$\boldsymbol{p}_0 = (\theta_0 - \theta_1)z_0 + \theta_0,$$

$$\boldsymbol{p}_1 = (\theta_1 - \theta_0)z_1 + \theta_1.$$

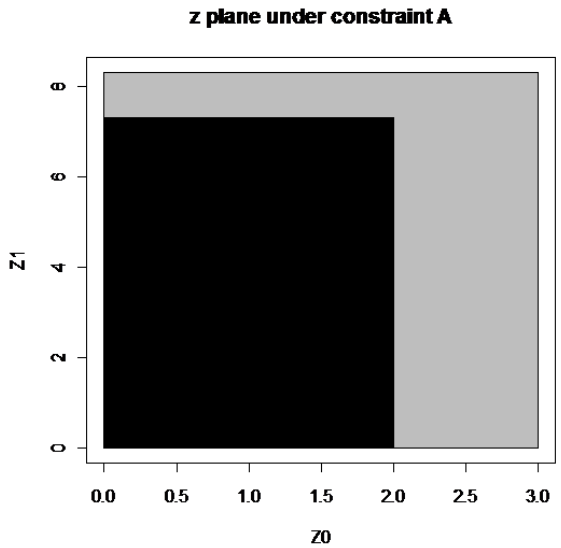
Constraint A

Setting

$\bar{z}_0(\theta) = min(\frac{\theta_{02}}{\theta_{12}-\theta_{02}}, \frac{\theta_{01}}{\theta_{11}-\theta_{01}})$ and $\bar{z}_1(\theta) = min(\frac{\theta_{10}}{\theta_{00}-\theta_{10}}, \frac{\theta_{11}}{\theta_{01}-\theta_{11}})$.

The identification region in the $\boldsymbol{z}$ plane is:

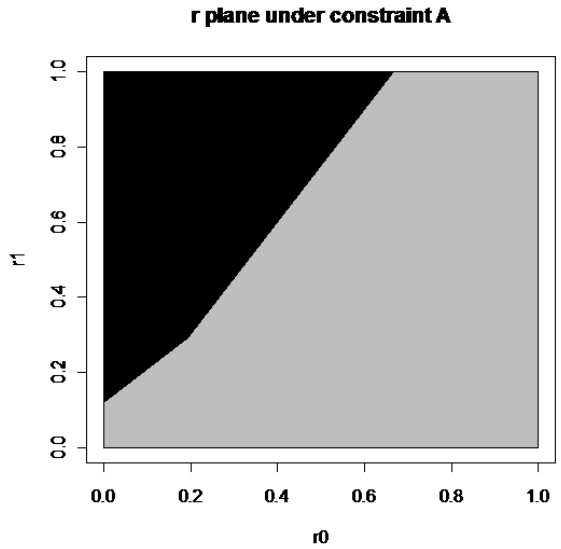
$$\{(z_0, z_1) : 0 < z_0 < \bar{z}_0(\theta),\ \ 0 < z_1 < \bar{z}_1(\theta)\}.$$



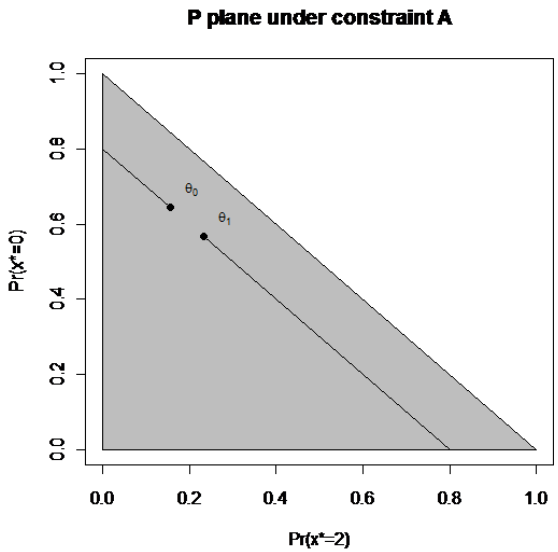
Constraint A

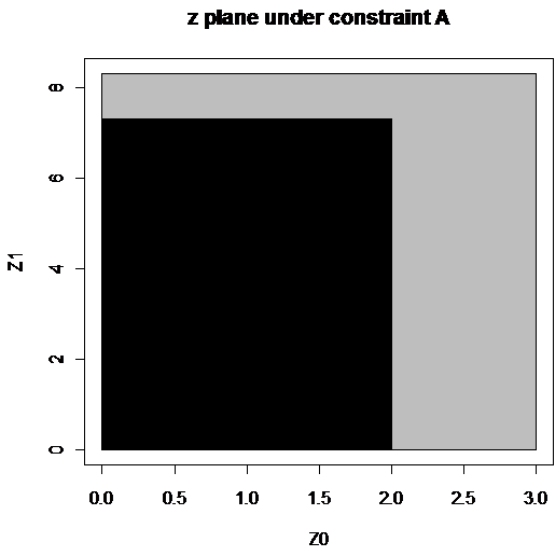
The identification region in $\boldsymbol{r}$ plane is:

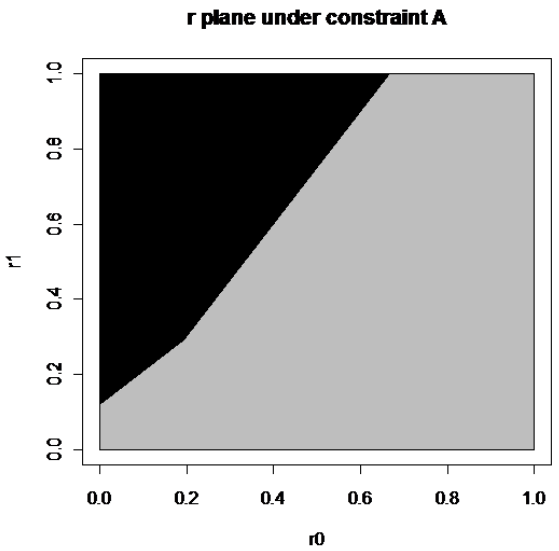
$$\mathbb{Q}_A(\boldsymbol{\theta}) = \left\{ (r_0, r_1) : r_1 > \frac{\bar{z}_0(\boldsymbol{\theta}) + 1}{\bar{z}_0(\boldsymbol{\theta})} r_0, r_1 > \frac{\bar{z}_1(\boldsymbol{\theta})}{\bar{z}_1(\boldsymbol{\theta}) + 1} r_0 + \frac{1}{\bar{z}_1(\boldsymbol{\theta}) + 1}, \right.$$
$$\left. 0 < r_0, r_1 < 1 \right\}.$$



Constraint A

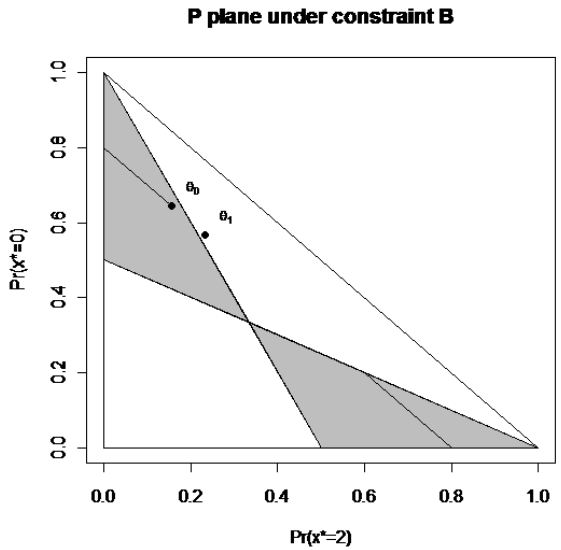






Constraint B

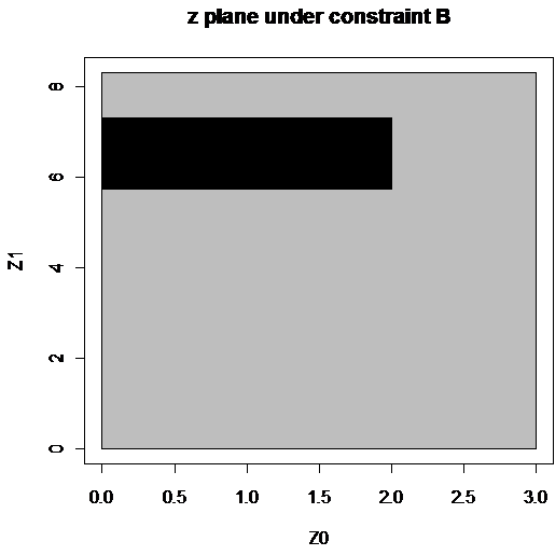
Under constraint B, there are four more restrictions:
$p_{00} > p_{01}$, $p_{01} > p_{02}$, $p_{10} < p_{11}$, and $p_{11} < p_{12}$.
The identification region in the $\boldsymbol{p}$ plane under the same setting as under constraint A:

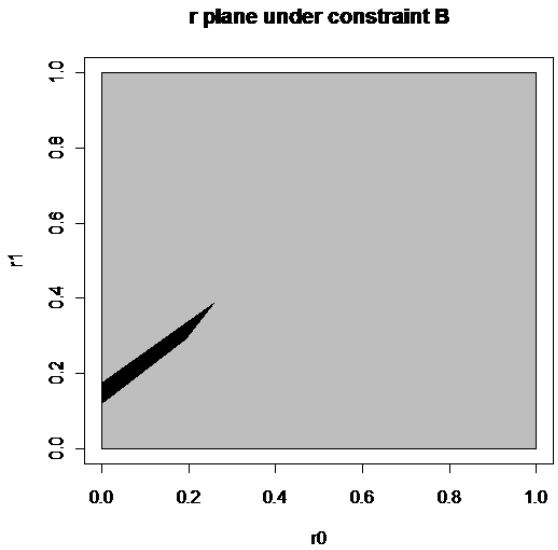


- The identification region $\mathbb{Q}_B(\theta)$ may be empty.

Constraint B

When θ is compatible:
using the same processes as for constraint A, the identification region in the ***z*** plane and the identification region in the ***r*** plane:





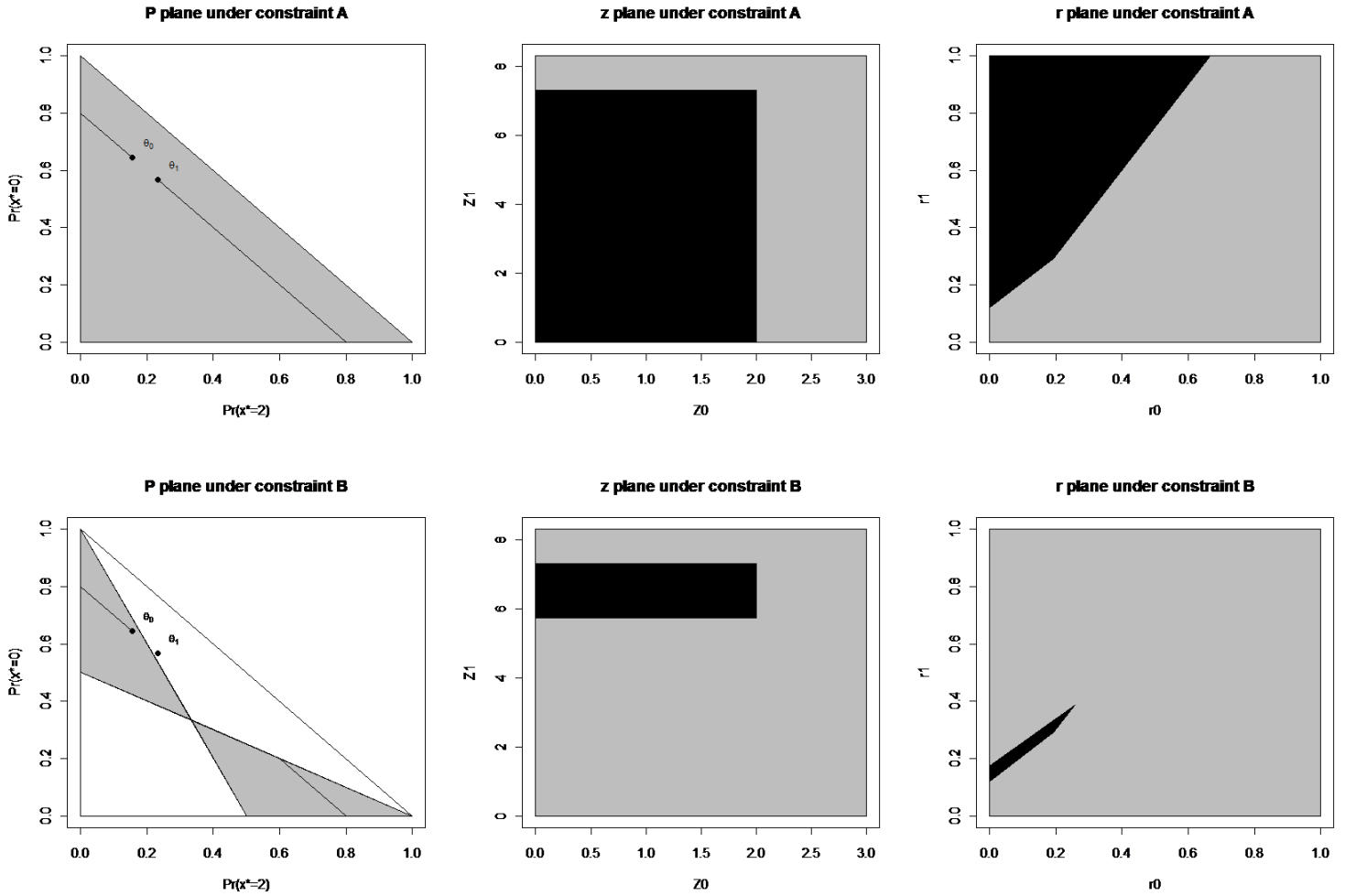
General Results

Theorem 1

If θ is compatible with constraint B:

- If $\boldsymbol{\theta} \in \mathbb{P}_B$, then $\mathbb{Q}_A(\boldsymbol{\theta}) = \mathbb{Q}_B(\boldsymbol{\theta})$. Otherwise, $\mathbb{Q}_B(\boldsymbol{\theta}) \subset \mathbb{Q}_A(\boldsymbol{\theta})$.
- Constraint A yields an infinite upper bound on the odds ratio.
  Constraint B yields a finite upper bound on the odds ratio iff θ_0 is outside the prior region for $\boldsymbol{p_0}$ and θ_1 is outside the prior region for $\boldsymbol{p_1}$.
- Constraints A and B yield the same lower bound on the odds ratio.

Comparison Between Constraint A and B



Collapsing Exposure to Two Categories

To compare with, let's consider the identification region obtained from the normal 2×2 table.
The apparent exposure status is:

$$X^{**} = \begin{cases} 0 & \text{if } X^* \in \{0, 1\}, \\ 1 & \text{if } X^* = 2, \end{cases}$$

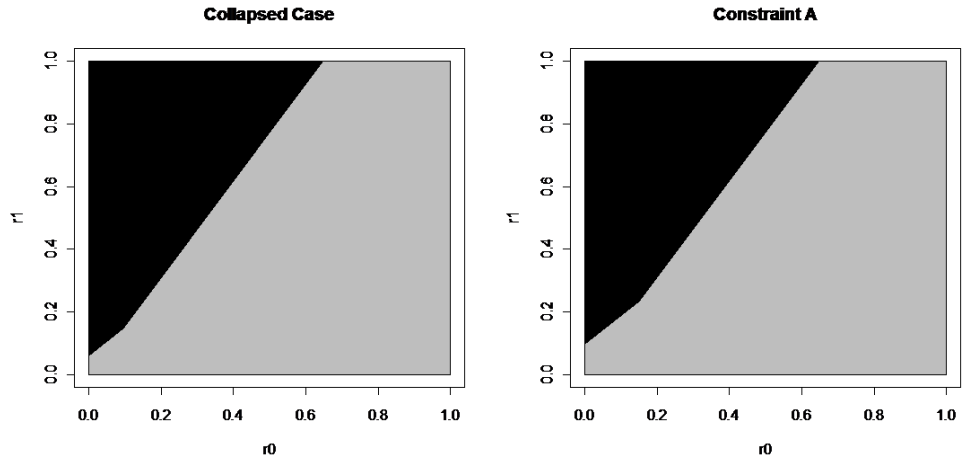
Misclassification can be described by:

- specificity: $p_0^* = Pr\{X^{**} = 0 \mid X = 0\} = p_{00} + p_{01}$,
- sensitivity: $p_1^* = Pr\{X^{**} = 1 \mid X = 1\} = p_{12}$.

Collapsing Exposure to Two Categories

$$\bar{z}_0^* = \frac{\theta_{02}}{\theta_{12}-\theta_{02}} \geq \bar{z}_0; \bar{z}_1^* = \frac{1-\theta_{12}}{\theta_{12}-\theta_{02}} \geq \bar{z}_1.$$

The identification region in the ***r*** plane between collapsed case and constraint A:



General Results

Theorem 2

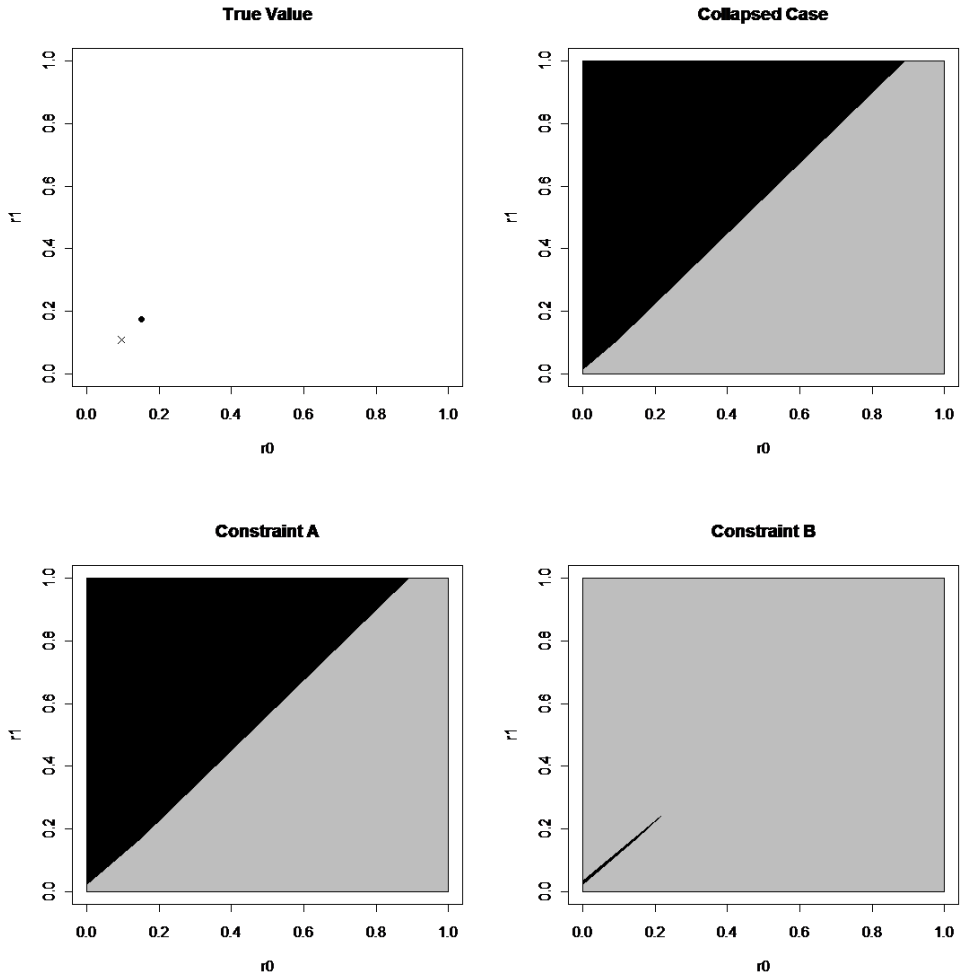
- $\mathbb{Q}_A(\boldsymbol{\theta}) \subseteq \mathbb{Q}^*(\boldsymbol{\theta}^*)$.
- The collapsed case yields an infinite upper bound on the odds ratio.
- The lower bound on the odds ratio in the collapsed case is not larger than under constraint A.

Examples

For a given $r_0 (r_0 = 0.05)$ and classification rules $(\boldsymbol{p}_0 = (0.900, 0.075, 0.025)$ and $\boldsymbol{p}_1 = (0.200, 0.300, 0.500))$, let's compare the situations for different OR.

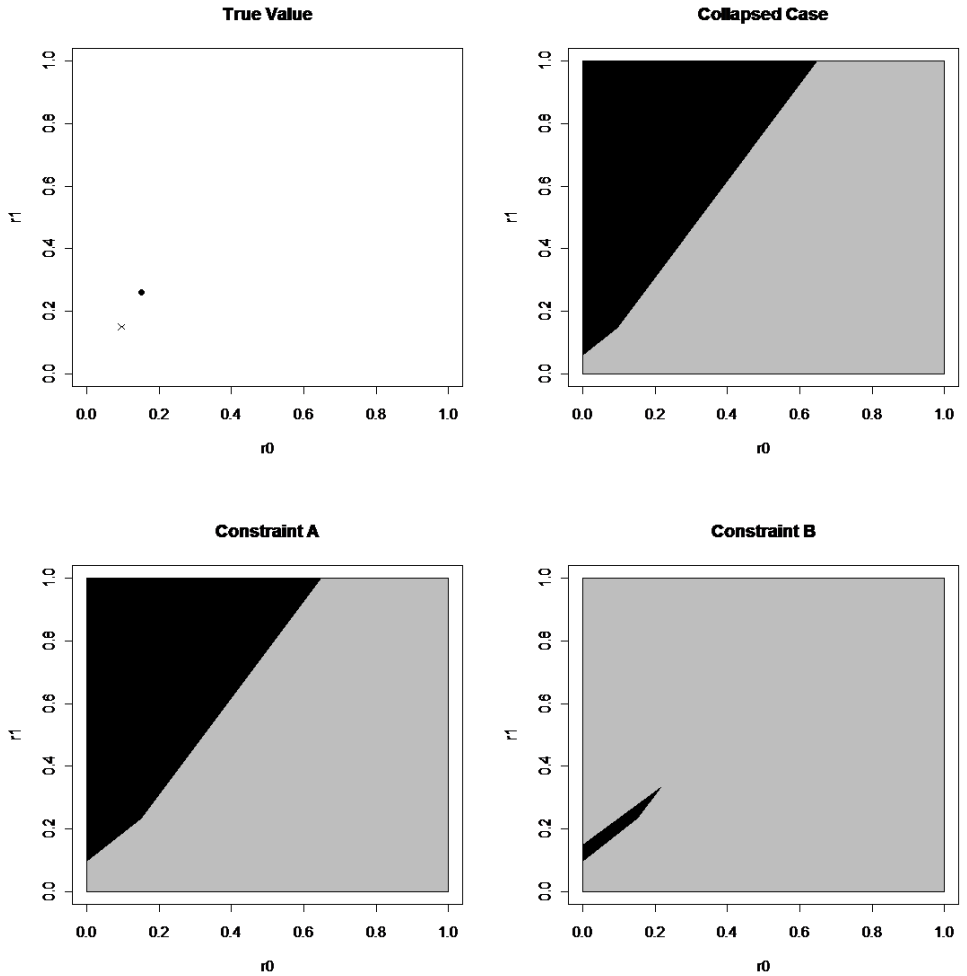
Examples

$OR = 1.2$



Examples

$OR = 2.0$



Limiting Posterior Distribution

The joint posterior density of all the parameters given the data is:

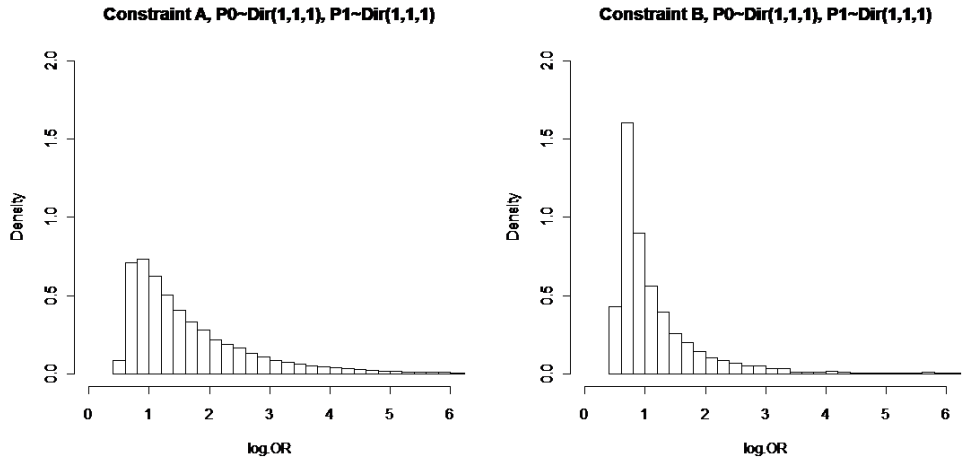
$$f(r_0, r_1, \theta_0, \theta_1 \mid X^*, Y) = f(\theta_0, \theta_1 \mid X^*, Y) f(r_0, r_1 \mid \theta_0, \theta_1).$$

As the sample size increases:

- $f(\theta_0, \theta_1 \mid X^*, Y)$ will converge to a point mass at the true values of $\boldsymbol{\theta_0}$ and $\boldsymbol{\theta_1}$;
- $f(r_0, r_1 \mid \theta_0, \theta_1)$ is proportional to $f(r_0, r_1, \boldsymbol{\theta_0}, \boldsymbol{\theta_1})$.

Examples

Assume $p_{00}, p_{01}, p_{02} \sim Dirichlet(1,1,1)$ and $p_{10}, p_{11}, p_{12} \sim Dirichlet(1,1,1)$ with the additional truncation to the assumed prior region $\mathbb{P}$.
Use the setting of $r_0 = 0.5$, $OR = 2$, $\boldsymbol{p}_0 = (0.900, 0.075, 0.025)$ and $\boldsymbol{p}_1 = (0.200, 0.300, 0.500)$.



Thanks!