Identification in Multivariate Partial Observability Probit

Dale J. Poirier
University of California, Irvine, USA
September 15, 2011

Abstract

Poirier (1980, JoE) considered a bivariate probit model in which the binary dependent variables $y_1$ and $y_2$ were not observed individually, but the product $z = y_1 \cdot y_2$ was observed. This paper expands this notion of partial observability to multivariate settings.
Bivariate Probit

• Consider $N$ independent observations from the \textit{latent bivariate regression model}

\[
\begin{bmatrix}
Y_{n1}^* \\
Y_{n2}^*
\end{bmatrix} \mid X_1, X_2, \theta \sim \mathcal{N}_2\left(\begin{bmatrix} x_n' \beta_1 \\ x_n' \beta_2 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right) \quad (n = 1, 2, \ldots, N),
\]

where $\theta = [\beta_1', \beta_2', \rho]'$ is an unknown parameter vector.
The **bivariate probit (BP) model** arises when only the sign of $Y_{ni}^*$ is observed, i.e., $Y_{ni} = 1$ if $Y_{ni}^* > 0$, and $Y_{ni} = 0$ if $Y_{ni}^* \leq 0$ ($i = 1, 2$).


- Chib and Greenberg (1998, *Biometrika*) provided a Bayesian analysis of **multivariate probit (MP)**.

The **bivariate ordered probit model** arises when $Y_{ni}^*$ ($i = 1, 2$) are observed in more than two categories.
• Of primary concern here is extending the case of *bivariate partial observability (BPO) probit* introduced in Poirier (1980, *JoE*) in which only \( Z_n = Y_{n1} \cdot Y_{n2} \) is observed. Also considered BPO in sample selections models (Heckit-like estimators).

• BPO can be thought of as a 2×2 contingency table with covariates and partial observability as suggested below.

<table>
<thead>
<tr>
<th>Cell Counts</th>
<th>( y_1 = 0 )</th>
<th>( y_1 = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y_2 = 0 )</td>
<td>( n_{00} )</td>
<td>( n_{10} )</td>
</tr>
<tr>
<td>( y_2 = 1 )</td>
<td>( n_{01} )</td>
<td>( n_{11} )</td>
</tr>
</tbody>
</table>

- In BPO, we only observe \( (n_{01} + n_{10} + n_{01}) = N - n_{11} \) and \( n_{11} \).
**Example:** Consider a two-agent committee that requires both agents to be in favor in order for a motion to pass.

- For example, consider a man and a women who contemplate getting married in the current year.

- You observe $Z = 1$ if they get married, and $Z = 0$ if they don’t.

- But you don’t observe their individual decisions $Y_1 = 0, 1$ and $Y_2 = 0, 1$. 

Note: There are analogs of partial observability in the analysis of contingency tables in which cells cannot be fully distinguished.


- In both cases issues of identifiability require careful analysis.

- An important difference between the these contingency table approach and BPO is the introduction of covariates provides ties across cells beyond the requirement that cell probabilities must add to unity.
## Table 1: Applications with Partial Observability

<table>
<thead>
<tr>
<th>Field</th>
<th>Article</th>
<th>Discrete Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>agriculture</td>
<td>Dimara, and Skuras (2003)</td>
<td>producer aware of innovation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>producer adopts innovation</td>
</tr>
<tr>
<td>banking</td>
<td>Dwyer and Hassan (2007)</td>
<td>bank fails</td>
</tr>
<tr>
<td></td>
<td></td>
<td>bank suspends payments</td>
</tr>
<tr>
<td>credit</td>
<td>Swain (2002)</td>
<td>bank’s decision on access</td>
</tr>
<tr>
<td></td>
<td></td>
<td>household’s demand for loans</td>
</tr>
<tr>
<td>education</td>
<td>Ballou (1996)</td>
<td>individual seeks teaching job</td>
</tr>
<tr>
<td></td>
<td></td>
<td>educator offers a job</td>
</tr>
<tr>
<td></td>
<td></td>
<td>working</td>
</tr>
<tr>
<td>development</td>
<td>Glewwe (1996)</td>
<td>private vs. public sector</td>
</tr>
<tr>
<td>IMF</td>
<td>Przeworski and Vreeland (2002);</td>
<td>decision by IMF to extend financing</td>
</tr>
<tr>
<td></td>
<td>Rajbhandari (2011)</td>
<td>decision by a gov. to apply for assistance</td>
</tr>
<tr>
<td>immigration</td>
<td>Aydemir (2002)</td>
<td>individual applies to emigrate</td>
</tr>
<tr>
<td></td>
<td></td>
<td>host decides whether to accept</td>
</tr>
<tr>
<td>foreign</td>
<td>Konig (2003)</td>
<td>ownership advantages hold</td>
</tr>
<tr>
<td>investment</td>
<td></td>
<td>location advantages hold</td>
</tr>
<tr>
<td></td>
<td></td>
<td>internalization advantages hold</td>
</tr>
<tr>
<td>labor</td>
<td>Mohanty (1992)</td>
<td>individual seeks employment</td>
</tr>
<tr>
<td></td>
<td></td>
<td>firm selects individual</td>
</tr>
<tr>
<td>law and</td>
<td>Feinstein (1990)</td>
<td>violation</td>
</tr>
<tr>
<td>economics</td>
<td></td>
<td>detection</td>
</tr>
<tr>
<td></td>
<td></td>
<td>richer school district</td>
</tr>
<tr>
<td>mergers</td>
<td>Brasington (2003)</td>
<td>poorer school district</td>
</tr>
</tbody>
</table>
Farber (1982, *Econometrica*) considered an intermediate case between BP and POBP which can be visualized as the $2 \times 2$ contingency table with covariates and the following structure (*censored probit*).

### Cell Counts

<table>
<thead>
<tr>
<th></th>
<th>$y_1 = 0$</th>
<th>$y_1 = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_2 = 0$</td>
<td>$n_{00}$</td>
<td>$n_{10}$</td>
</tr>
<tr>
<td>$y_2 = 1$</td>
<td>$n_{01}$</td>
<td>$n_{11}$</td>
</tr>
</tbody>
</table>

- In this case, we observe $(n_{00} + n_{01})$, $n_{10}$, and $n_{11}$. 
• Maximum likelihood estimation of BPO is included in STATA.

• Cavanagh and Sherman (1998, *JoE*) introduced a class of rank estimators of scaled coefficients in semiparametric monotonic linear index models.
  
  ○ CS also considered *single equation multiple indices models* including BPO probit.
  
  ○ Importantly, their results imply semiparametric analysis of partial observability models without the normality assumption are identified.
Let $\Phi(\cdot)$ denote the univariate standard normal cdf, and $\Phi_2(0, 0; \rho)$ denote the bivariate standard normal cdf with correlation $\rho$.

The likelihood functions for BP and BPO are

$$
\mathcal{L}^{BP}(\theta; y) = \prod_{n=1}^{N} \left[ P_{Y_n}([1, 1]' | \theta) \right]^{y_{n1}y_{n2}} \left[ P_{Y_n}([1, 0]' | \theta) \right]^{y_{n1}(1 - y_{n2})} 
\cdot \left[ P_{Y_n}([0, 1]' | \theta) \right]^{(1 - y_{n1})y_{n2}} \left[ P_{Y_n}([0, 0]' | \theta) \right]^{(1 - y_{n1})(1 - y_{n2})},
$$

$$
\mathcal{L}^{BPO}(\theta; z) = \prod_{n=1}^{N} \left[ P_{Y_n}([1, 1]' | \theta) \right]^{z_n} \left[ 1 - P_{Y_n}([1, 1]' | \theta) \right]^{1 - z_n},
$$

where $y = [y_{11}, \ldots, y_{N1}, y_{12}, \ldots, y_{N2}]'$ and $z = [z_1, z_2, \ldots, z_N]'$ are observed and the component probabilities are given in the following table.
Joint Probabilities $P_{Y_n}([i,j]^\prime | \theta)$ (i, j = 0, 1) for BP

<table>
<thead>
<tr>
<th>$y_{n1} = 0$</th>
<th>$y_{n1} = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_{n2} = 0$</td>
<td>$1 - \Phi(x_n^\prime \beta_2) - \Phi(x_n^\prime \beta_1)$</td>
</tr>
<tr>
<td></td>
<td>$+ \Phi_2(x_n^\prime \beta_1, x_n^\prime \beta_2, \rho)$</td>
</tr>
<tr>
<td>$y_{n2} = 1$</td>
<td>$\Phi(x_n^\prime \beta_2) - \Phi_2(x_n^\prime \beta_1, x_n^\prime \beta_2, \rho)$</td>
</tr>
</tbody>
</table>
Meng and Schmidt (1983, *IER*) compared the relative variances for BP vs. BPO for a small number of parameter settings.

- The information loss under BPO can be "surprisingly large," particularly near points that are not identified.

- The cost increases with the fraction of the data for which the dependant variable is imperfectly observed.
• If $\rho = 0$:
  ○ BP reduces to two univariate probit models with a block diagonal information matrix.
  ○ The BPO information matrix is \textit{not} block diagonal.

• If $\rho \neq 0$:
  ○ BP is more efficient than univariate probit.
  ○ In contrast, the underlying latent system shows no gains from pooling the two equations unless restrictions are added.
    ▶ The nonlinearity of the model is why the OLS efficiency result for the linear latent model does not carry over.
    ▶ But this result does carry over if a \textit{bivariate linear probability model} is estimated.
Multivariate Probit

- Consider a sample of $N$ independent observations from a multivariate probit model with latent variable representation

$$Y_{n1}, ..., Y_{nJ} | x_1, ..., x_N, \beta, \rho \sim \mathcal{N}_J(\mu_n, \Omega(\rho))$$

where

$$\mu_n = \begin{bmatrix} \mu_{n1} \\ \vdots \\ \mu_{nJ} \end{bmatrix} = \begin{bmatrix} x_n' \beta_1 \\ \vdots \\ x_n' \beta_J \end{bmatrix} = (I_J \otimes x_n') \beta \quad (n = 1, ..., N), \quad (1)$$

$$\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_J \end{bmatrix}, \quad \Omega(\rho) = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1J} \\ \rho_{12} & 1 & \cdots & \rho_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1J} & \rho_{2J} & \cdots & 1 \end{bmatrix}, \quad \rho = \begin{bmatrix} \rho_{12} \\ \rho_{13} \\ \vdots \\ \rho_{(J-1)J} \end{bmatrix}, \quad (2)$$
and you only observe the binary outcomes

\[
Y_{nj} = \begin{cases} 
1, & \text{if } Y_{nj}^* > 0 \\
0, & \text{if } Y_{nj}^* \leq 0 
\end{cases}, \quad Y_n = [Y_{n1}, ..., Y_{nJ}]'.
\]  

- Analysis here is conditional on the $K \times 1$ vectors $x_n$ ($n = 1, 2, ..., N$).
- Stack all parameters into the $M = 3K + 3$ vector $\theta = [\beta', \rho']'$.
- Finally, let $\phi_j(\cdot|m, \Sigma)$ denote a $J$-dimensional normal pdf with mean $m$ and covariance matrix $\Sigma$ and denote the corresponding cdf by $\Phi_j(\cdot|m, \Sigma)$.
The *multivariate probit (MP)* choice probability is

\[
P_{Y_n}(y_n | \theta) = \text{Prob}(Y_n = y_n | \theta) = \int_{\Psi_{nJ}} \ldots \int_{\Psi_{n1}} \Phi_J(y^*_n | (I_J \otimes x_n') \beta, \Omega(\rho)) \; dy^*_n1 \ldots dy^*_nJ, \tag{4}
\]

where the regions of integration are given by

\[
\Psi_{nj} = \begin{cases} 
  (-\infty, 0], & \text{if } y_{nj} = 0 \\
  (0, \infty), & \text{if } y_{nj} = 1 
\end{cases} \quad (n = 1, \ldots, N; \ j = 1, \ldots, J). \tag{5}
\]
• A convenient representation for MP choice probability $P_{Y_n}(y_n|\theta)$ can be obtained by reparameterizing $\mu_n$ and $\rho$ to

$$\tilde{\mu}_n = [q_{n1}\mu_{n1}, ..., q_{nJ}\mu_{nJ}]',$$

$$\tilde{\rho}_n = [\tilde{\rho}_{n12}, ..., \tilde{\rho}_{(J-1)J}]' = [q_{n1}q_{n2}\rho_{12}, ..., q_{n(J-1)}q_{nJ}\rho_{(J-1)J}]',$$

where $q_{nj} = 2y_{nj} - 1$ ($n = 1, ..., N; j = 1, ..., J$). Then

$$P_{Y_n}(y_n|\theta) = \Phi_J(\tilde{\mu}_n|0_J, \Omega(\tilde{\rho}_n)).$$ (6)

• The log-likelihood function for J-dimensional multivariate probit is

$$L^{MP}(\theta; y) = \sum_{n=1}^{N} \ln \left[ \Phi_J(\tilde{\mu}_n|0_J, \Omega(\tilde{\rho}_n)) \right].$$ (7)
Barrenechea (2007) derived
\[
\frac{\partial L^{\text{MP}}(\theta; y)}{\partial \beta_j} = \sum_{n=1}^{N} \left[ \frac{q_{nj} \phi(q_{nj} x_n' \beta_j) \Phi_{J-1}(\tilde{\mu}_{n(-j)} | \lambda_n, \Lambda_n)}{\Phi_J(\tilde{\mu}_n | 0_J, \Omega(\tilde{\rho}_n))} \right] x_n (j = 1, ..., J), \quad (8)
\]

\[
\frac{\partial L^{\text{MP}}(\theta; y)}{\partial \rho_{ij}} = \sum_{n=1}^{N} \frac{q_{ni} q_{nj} \phi_2(\tilde{\mu}_n, \tilde{\mu}_n | 0_2, A_{nij}) \Phi_{J-2}(\tilde{\mu}_{n(-ij)} | \tilde{b}_{nij}, \tilde{C}_{nij})}{\Phi_J(\tilde{\mu}_n | 0_J, \Omega(\tilde{\rho}_n))} (i, j = 1, ..., J; i < j), \quad (9)
\]

\( \Omega_{-j}(\tilde{\rho}) \) is the \((J - 1) \times (J - 1)\) matrix obtained by omitting the \(j\)th row and column of \(\Omega(\tilde{\rho})\), \( \Omega_{-ij}(\tilde{\rho}) \) is the \((J - 2) \times (J - 2)\) matrix obtained by omitting the \(i\)th and \(j\)th rows and columns of \(\Omega(\tilde{\rho})\),

\[
\tilde{\omega}_{-j} = [\tilde{\rho}_{1j}, ..., \tilde{\rho}_{(j-1)j}, \tilde{\rho}_{(j+1)j}, ..., \tilde{\rho}_{Jj}]',
\]

\[
\tilde{\mu}_{n(-j)} = [\tilde{\mu}_{1j}, ..., \tilde{\mu}_{n(j-1)}, \tilde{\mu}_{n(j+1)}, ..., \tilde{\mu}_{nJ}]',
\]
\[ \tilde{\mu}_{n(-ij)} = [\tilde{\mu}_{n1}, \ldots, \tilde{\mu}_{n(i-1)}, \tilde{\mu}_{n(i+1)}, \ldots, \tilde{\mu}_{n(j-1)}, \tilde{\mu}_{n(j+1)}, \ldots, \tilde{\mu}_{jj}]', \]

\[ A_{nij} = \begin{bmatrix} 1 & \tilde{\rho}_{nij} \\ \tilde{\rho}_{nij} & 1 \end{bmatrix}, \]

\[ W_{nij} = \begin{bmatrix} \omega_{ni} & \omega_{nj} \end{bmatrix}, \]

\[ \tilde{C}_{nij} = \tilde{\Omega}_{nij} - W_{nij} A_{nij} W_{nij}', \]

\[ \tilde{b}_{nij} = \tilde{W}_{nij} A^{-1}_{nij} \begin{bmatrix} \tilde{\mu}_{ni} \\ \tilde{\mu}_{nj} \end{bmatrix}. \]

\[ \lambda_n = (q_{nj} x_n' \beta_j) \tilde{\omega}_{-j}, \]

\[ \Lambda_n = \Omega_{-j}(\tilde{\rho}) - \tilde{\omega}_{-j} \tilde{\omega}_{-j}'. \]
Table 2 expresses

\[ P_Y(y_n | \theta) = \Phi_J(\tilde{\mu}_n | 0_J, \Omega_J(\tilde{\rho}_n)) \]  \hspace{1cm} (6)

in terms of the original parameters \( \mu_n \) and \( \rho \) in the case \( J = 3 \).

<table>
<thead>
<tr>
<th>( y_{n3} = 0 )</th>
<th>( y_{n2} )</th>
<th>( y_{n1} = 0 )</th>
<th>( y_{n1} = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y_{n2} = 0 )</td>
<td>( \Phi_3(-\mu_{n1}, -\mu_{n2}, -\mu_{n3}, \rho_{12}, \rho_{13}, \rho_{23}) )</td>
<td>( \Phi_3(\mu_{n1}, -\mu_{n2}, -\mu_{n3}, -\rho_{12}, -\rho_{13}, \rho_{23}) )</td>
<td></td>
</tr>
<tr>
<td>( y_{n2} = 1 )</td>
<td>( \Phi_3(-\mu_{n1}, \mu_{n2}, -\mu_{n3}, -\rho_{12}, \rho_{13}, -\rho_{23}) )</td>
<td>( \Phi_3(\mu_{n1}, \mu_{n2}, -\mu_{n3}, \rho_{12}, -\rho_{13}, -\rho_{23}) )</td>
<td></td>
</tr>
<tr>
<td>( y_{n3} = 1 )</td>
<td>( y_{n2} = 0 )</td>
<td>( \Phi_3(-\mu_{n1}, -\mu_{n2}, \mu_{n3}, \rho_{12}, -\rho_{13}, -\rho_{23}) )</td>
<td>( \Phi_3(\mu_{n1}, -\mu_{n2}, \mu_{n3}, -\rho_{12}, \rho_{13}, -\rho_{23}) )</td>
</tr>
<tr>
<td>( y_{n2} = 1 )</td>
<td>( \Phi_3(-\mu_{n1}, \mu_{n2}, \mu_{n3}, -\rho_{12}, -\rho_{13}, \rho_{23}) )</td>
<td>( \Phi_3(\mu_{n1}, \mu_{n2}, \mu_{n3}, \rho_{12}, \rho_{13}, \rho_{23}) )</td>
<td></td>
</tr>
</tbody>
</table>
The information matrix for multivariate probit is

\[ I_{MP}(\theta) = E_Y \left( \frac{\partial L_{MP}(\theta; y)}{\partial \theta} \right) \left( \frac{\partial L_{MP}(\theta; y)}{\partial \theta'} \right) \]

\[ = \sum_{n=1}^{N} \left( \sum_{i_1=0}^{1} ... \sum_{i_J=0}^{1} g_n(i|\theta) \right), \tag{14} \]

where the expected information in cell \( i = [i_1, ..., i_J]' \) is

\[ g_n(i|\theta) = \frac{1}{P_Y(y_n=i|\theta)} \left[ \frac{\partial P_Y(y_n=i|\theta)}{\partial \theta} \right] \left[ \frac{\partial P_Y(y_n=i|\theta)}{\partial \theta'} \right]. \tag{15} \]
Multivariate Partial Observability Probit

- A direct extension of Poirier (1980) to multivariate partial observability (MPO) probit corresponds to observing only \( Z_n = \prod_{j=1}^{J} Y_{nj} \) (\( n = 1, ..., N \)).

**Table 3: Sampling Distribution for MPO**

\[
\text{Prob}(z_n = i | \theta) \quad (i = 0, 1)
\]

\[
\begin{array}{c|c}
  z_n = 0 & 1 - \Phi_J(\mu_n; \rho) \\
  z_n = 1 & \Phi_J(\mu_n; \rho)
\end{array}
\]
The log-likelihood for MPO is

$$L^{\text{MPO}}(\theta; z) = \sum_{n=1}^{N} (1 - z_n) \ln[1 - P_z(z_n = 1|\theta)] + z_n \ln[P_z(z_n = 1|\theta)]$$

$$= \sum_{n=1}^{N} (1 - z_n) \ln[1 - \Phi_j(\mu; \rho)] + z_n \ln[\Phi_j(\mu_n; \rho)],$$

where $z = [z_1, ..., z_N]'$ is observed.
The information matrix for MPO is

\[ I_{\text{MPO}}(\theta) = \mathbb{E}_Z \left( \frac{\partial L_{\text{MPO}}(\theta; z)}{\partial \theta} \right) \left( \frac{\partial L_{\text{MPO}}(\theta; z)}{\partial \theta'} \right) \]

\[ = \sum_{n=1}^{N} \sum_{i=0}^{1} h_n(i|\theta) \]  

\[ = \sum_{n=1}^{N} \frac{1}{\Phi_j(\mu_n; \rho)[1 - \Phi_j(\mu_n; \rho)]} \left[ \frac{\partial \Phi_j(\mu_n; \rho)}{\partial \theta} \right] \left[ \frac{\partial \Phi_j(\mu_n; \theta)}{\partial \theta'} \right], \]  

where

\[ h_n(i|\theta) = \frac{1}{P_Z(z_n = i|\theta)} \left[ \frac{\partial P_Z(z_n = i|\theta)}{\partial \theta} \right] \left[ \frac{\partial P_Z(z_n = i|\theta)}{\partial \theta'} \right] \quad (i = 0, 1). \]
• The additional information in MP over MPO is

\[ I^{\text{MP}}(\theta) - I^{\text{MPO}}(\theta) = \]

\[
\sum_{n=1}^{N} \frac{1}{1 - P_Y(y_n = \iota_j | \theta)} \sum_{i=0}^{1} \cdots \sum_{i=0}^{1} \left[ 1 - P_Y(y_n = \iota_j | \theta) - P_Y(y_n = i | \theta) \right] g_n(i | \theta) \]  (18)

where \( i = [i_1, \ldots, i_J]' \), \( \iota_J = [1, \ldots, 1]' \), and \( g_n(i | \theta) \) is given by (15).

○ Clearly (18) is positive semidefinite.

○ The internal sum in (18) is over the cells for \( y_n \) that cannot be distinguished when \( z_n = 0 \).
• MPO reflects unanimous consent among J agents.
  
  ◦ J = 2 (marriage)
  
  ◦ J = 12 (American criminal justice)
    
    ▶ Z_n = 1 (guilty)
    
    ▶ Z_n = 0 implies unanimous agreement on “not guilty” or a hung jury.
• The next section discusses conditions under which (16) is positive definite, and hence, \( \theta \) is locally identified.

  ◦ But even when \( \theta \) is identified, MPO is unlikely to provide much information on \( \theta \) when \( J > 2 \).

  ◦ So it is natural to look at other types of partial information that involve *more* information than MPO, but *less* than MP.
Suppose the complete set of $J(J - 1)/2$ bivariate products

\[ Z_{nij} = Y_{ni} \cdot Y_{nj} \quad (i, j = 1, ..., J; i < j) \]

are observed. BPO applied to all possible pairs in MP is defined to be as multivariate bivariate partial observability (MBPO).
○ MP uncovers $2^J$ cells for $y_n$.

○ MPO uncovers one cell, $y_n = [1, \ldots, 1]'$, but cannot distinguish among the other $2^J - 1$ cells.

○ MBPO uncovers $2^J - J - 1$ cells, but cannot distinguish among the $J + 1$ cells with $z_{nij} = 0$ (i.e., the cell $y_n = 0_j$ and $J$ cells with exactly one $y_{nij} = 1$).

○ As $J$ increases the proportion of cells that can be distinguished increases.
• How interesting is the MBPO data situation?

○ A data collecting agency can use MBPO to design data releases that are only partially informative.

○ Consider asking J binary (yes/no) questions including a highly sensitive one in which \( y_{nj} = 0 \) is embarrassing.
  ▶ If individuals are only asked whether all their answers equal one (MPO), or whether all answers on pairs of equations are one (MBPO), then a “no” answer does not imply \( y_{nj} = 0 \).

○ Other suggestions?
• The marginal distribution of $Z_{nij}$ in MBPO is Bernoulli with

$$P_{Z_n}(Z_{nij} = 1 | \theta) = \text{Prob}(Z_{nij} = 1 | \theta)$$

$$= \text{Prob}(Y_{ni} = 1, Y_{nj} = 1 | \theta)$$

$$= \Phi_2(\mu_{ni}, \mu_{nj} | 0_2, \Omega(\rho_{ij})).$$

• Poirier (1980) gave conditions such that $Z_{nij}$ identifies $\beta_i$, $\beta_j$, and $\rho_{ij}$.

• The joint sampling distribution $Z_{nij} = Y_{ni} \cdot Y_{nj}$ ($i, j = 1, \ldots, J ; i < j$) provides a likelihood function for identifying $\theta$. 
Unlike using all $\binom{J}{2}$ possible BPs to estimate MP parameters [e.g., Kimhi (1994, *AJAE*)], MBPO uses the actual likelihood for $z_n = [z_{n12}, \ldots, z_{n1J}, z_{n23}, \ldots, z_{n(J-1)J}]'$, and *not* a quasi-likelihood.

“quasi” because the BPs are *not* independent.

MBPO involves J-dimensional integrals. BPO only requires two dimensional integrals.
• If $z_n$ uncovers a specific cell, let $d_n = 1$, and let $d_n = 0$ otherwise.

  - If $d_n = 1$, then $P_{Z}(Z_n | \theta) = \Phi_j(\tilde{\mu}_n; \tilde{\rho}_n)$.

  - If $d_n = 0$, then $\text{Prob}_{Z_n}(z_n = 0_{(J-1)J/2} | \theta)$ equals the sum of probabilities for all the cells that cannot be distinguished, i.e.,

$$P_{Z_n}(z_n = 0_{(J-1)J/2} | \theta) = P_Y(y_n = 0_J' | \theta) + P_Y(y_n = [1, 0, ..., 0]' | \theta)$$

$$+ ... + P_Y(y_n = [0, 0, ..., 1]' | \theta).$$  

\hfill (19)
The MBPO log-likelihood is

\[ L^{\text{MBPO}}(\theta; z) = \sum_{n=1}^{N} (1 - d_n) \ln \left[ \text{Prob}(Z_n = 0_{(J-1)J/2} | \theta) \right] + d_n \ln \left[ \Phi_J(\tilde{\mu}_n | 0_J, \Omega(\tilde{\rho}_n)) \right], \]

(20)

The information matrix for MBPO is

\[ I^{\text{MBPO}}(\theta) = \sum_{n=1}^{N} \frac{1 - d_n}{\text{Prob}(Z_n = 0_{(J-1)J/2} | \theta)} \left[ \frac{\partial \text{Prob}(Z_n = 0_{(J-1)J/2} | \theta)}{\partial \theta} \right] \left[ \frac{\partial \text{Prob}(Z_n = 0_{(J-1)J/2} | \theta)}{\partial \theta'} \right] + \frac{d_n}{\Phi_J(\tilde{\mu}_n | 0_J, \Omega_J(\tilde{\rho}_n))} \left[ \frac{\partial \Phi_J(\tilde{\mu}_n | 0_J, \Omega_J(\tilde{\rho}_n))}{\partial \theta} \right] \left[ \frac{\partial \Phi_J(\tilde{\mu}_n | 0_J, \Omega_J(\tilde{\rho}_n))}{\partial \theta'} \right]. \]

\[ I^{\text{MBPO}}(\theta) - I^{\text{MPO}}(\theta) \] and \[ I^{\text{MP}}(\theta) - I^{\text{MBPO}}(\theta) \] are positive semidefinite.
• Consider *trivariate partial observability (TPO)* \((J = 3)\) [Konig (2003, *Rev. World Econ*)].
The additional information in *trivariate probit (TP)* over TPO is

\[ I^{TP}(\theta) - I^{TPO}(\theta) = \]

\[
\sum_{n=1}^{N} \frac{1}{1 - P_{Yn}(i_3 | \theta)} \sum_{i_1=0}^{1} \sum_{i_2=0}^{1} \sum_{i_3=0}^{1} \left[ 1 - P_{Yn}(i_3 | \theta) - P_{Yn}(i | \theta) \right] g_n(i | \theta)
\]

where \( i = [i_1, i_2, i_3]' \) and

\[
g_n(i | \theta) = \frac{1}{P_{Y}(y_n = i | \theta)} \left[ \frac{\partial P_{Y}(y_n = i | \theta)}{\partial \theta} \right] \left[ \frac{\partial P_{Y}(y_n = i | \theta)}{\partial \theta'} \right]. \tag{15}
\]
Trivariate bivariate partial observability (TBPO) arises when only $z_{n12} = y_{n1} \cdot y_{n2}$, $z_{n13} = y_{n1} \cdot y_{n3}$, and $z_{n23} = y_{n2} \cdot y_{n3}$ are observed.
For some realizations $z_n = [z_{n12}, z_{n13}, z_{n23}]'$ of $Z_n = [Z_{n12}, Z_{n13}, Z_{n23}]'$, it is possible to recover the unobserved $y_n$.

- If $z_{n12} = z_{n23} = 1$, then $y_n = [1, 1, 1]'$.
- If $z_{n12} = 1$ and $z_{n23} = 0$, then $y_n = [1, 1, 0]'$.
- If $z_{n12} = 0$ and $z_{n23} = 1$, then $y_n = [0, 1, 1]'$.
- If $z_{n12} = z_{n23} = 0$ and $z_{n13} = 1$, then $y_n = [1, 0, 1]'$.
- Set $d_n = 1$ if any of these four cases occurs.
Otherwise, set $d_n = 0$ for the four remaining cells which cannot be distinguished when $z_{n12} = z_{n13} = z_{n23} = 0$ [see (19)]:

$$\text{Prob}(Z_n = 0_3 | \theta) = P_Y(y_n = 0_3 | \theta) + P_Y(y_n = [1, 0, 0]' | \theta) +$$

$$P_Y(y_n = [0, 1, 0]' | \theta) + P_Y(y_n = [0, 0, 1]' | \theta).$$  \hspace{1cm} (23)

See Table 4.
Table 4: Mapping $z_{12}$, $z_{23}$, $z_{13} \rightarrow y_n$ for TBPO

<table>
<thead>
<tr>
<th>$z_{n13}$ = 0</th>
<th>$z_{n13}$ = 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$z_{n12}$ = 0</td>
<td>$z_{n12}$ = 1</td>
</tr>
<tr>
<td>$z_{n23}$ = 0</td>
<td>[0, 0, 0]', [1, 0, 0]', [0, 1, 0]', [0, 0, 1]'</td>
</tr>
<tr>
<td>$z_{n23}$ = 1</td>
<td>[0, 1, 1]'</td>
</tr>
</tbody>
</table>

Note: Blank cells indicate impossible combinations.

- The *joint* distribution of $Z_n$, $P_Z(z_{n12}, z_{n13}, z_{n23} | \theta)$, is given in Table 5.
Table 5: Sampling Distribution for TBPO

| \(z_{n12}\) | \(z_{n13}\) | \(z_{n23}\) | \(P_{z_n}(z_{n12}, z_{n23}, z_{n13} | \theta)\) |
|---|---|---|---|
| 0 | 0 | 0 | \(1 - \Phi_3(\mu_{n1}, \mu_{n2}, -\mu_{n3}, \rho_{12}, -\rho_{13}, -\rho_{23})\) |
| 1 | 0 | 0 | \(\Phi_3(\mu_{n1}, \mu_{n2}, -\mu_{n3}, \rho_{12}, -\rho_{13}, -\rho_{23})\) |
| 0 | 1 | 0 | \(\Phi_3(\mu_{n1}, -\mu_{n2}, \mu_{n3}, -\rho_{12}, \rho_{13}, -\rho_{23})\) |
| 0 | 0 | 1 | \(\Phi_3(-\mu_{n1}, \mu_{n2}, \mu_{n3}, -\rho_{12}, -\rho_{13}, \rho_{23})\) |
| 1 | 1 | 1 | \(\Phi_3(\mu_{n1}, \mu_{n2}, \mu_{n3}, \rho_{12}, \rho_{13}, \rho_{23})\) |

**Note:** Other combinations of the \(z_{nij}\) are assigned probability zero.
The TBPO log-likelihood is

\[
L^{TBPO}(\theta; z) = \sum_{n=1}^{N} (1 - d_n) \ln [\text{Prob}(Z_n = 0, 3 | \theta)] \\
+ d_n \ln [\Phi_3 (\tilde{\mu}_n | 0, 3, \Omega_3(\tilde{\rho}_n))],
\]

where \( \theta = [\beta_1', \beta_2', \beta_3', \rho_{12}, \rho_{13}, \rho_{23}]' \) and

\[
I^{TBPO}(\theta) = \sum_{n=1}^{N} \frac{1}{P_{Z_n}([0, 0, 0]' | \theta)} \left[ \frac{\partial P_{Z_n}([0, 0, 0]' | \theta)}{\partial \theta} \right]^{-1} \left[ \frac{\partial P_{Z_n}([0, 0, 0]' | \theta)}{\partial \theta'} \right]^{-1} \\
+ g_n([1, 1, 0]' | \theta) + g_n([1, 0, 1]' | \theta) + g_n([0, 1, 1]' | \theta) + g_n([1, 1, 1]' | \theta)
\]
The information \textit{lost} in going from TBPO to TP is

\[
I^\text{TP}(\theta) - I^\text{TBPO}(\theta) = \sum_{n=1}^{N} \frac{1}{P_{Z_n}(0_3|\theta)} \times \\
\left( P_{Z_n}(0_3|\theta) - P_{Y_n}([0,0,0]'|\theta) \right) g_n([0,0,0]'|\theta) \\
+ \left( P_{Z_n}(0_3|\theta) - P_{Y_n}([1,0,0]'|\theta) \right) g_n([1,0,0]'|\theta) \\
+ \left( P_{Z_n}(0_3|\theta) - P_{Y_n}([0,1,0]'|\theta) \right) g_n([0,1,0]'|\theta) \\
+ \left( P_{Z_n}(0_3|\theta) - P_{Y_n}([0,0,1]'|\theta) \right) g_n([0,0,1]'|\theta)
\]

The information \textit{gained} in going from TPO to TBPO is

\[
I^\text{TBPO}(\theta) - I^\text{TPO}(\theta) = \sum_{n=1}^{N} g_n([1,1,0]'|\theta) + g_n([1,0,1]'|\theta) + g_n([0,1,1]'|\theta)
\]
• Just like in the case of MP, increasing $J$ by one adds new parameters, but also the possibility of more information on $\beta_j$ ($1 \leq j \leq J$) provided $\rho_{j(J+1)} \neq 0$. The same holds for partial observability.

○ Going from $J$ to $J+1$ increases the number of parameters by $K + J$. 
Identification

- Lindley (1971): “In passing it might be noted that unidentifiability causes no real difficulty in the Bayesian approach.” (assuming a *proper* prior)

- Kadane (1974): “… identification is a property of the likelihood function, and is the same whether considered classically or from the Bayesian approach.”

- Poirier (1988, *ET*): There is, however, no *Bayesian free lunch*. The “price” is that there exist quantities about which the data are uninformative, i.e., their marginal prior and posterior distributions are identical.
Consider the bivariate case $J = 2$.

The BPO likelihood is

$$
\mathcal{L}^{BPO}(\theta; z) = \prod_{n=1}^{N} \left[ P_{Y_n}([1, 1]'|\theta) \right]^{z_n} \left[ 1 - P_{Y_n}([1, 1]'|\theta) \right]^{1 - z_n},
$$

Because

$$
P_{Z_n}([1, 1]'|\theta) = \Phi_2(x_n'\beta_1, x_n'\beta_2, \rho_{12})
= \Phi_2(x_n'\beta_2, x_n'\beta_1, \rho_{12})
$$

there is a labeling problem.
  - If the function of interest is invariant to parameter permutation, then the labeling is not a problem.
    ▶ An example is prediction of $z$.
    ▶ However, obtaining posterior simulation convergence can be more challenging.
  - If the function of interest involves, say, only $\beta_1$, then the prior should also *not* be permutation invariant.
    ▶ Different marginal priors for $\beta_1$ and $\beta_2$ should be used.
    ▶ E.g., a dogmatic exclusion restriction on a component in $\beta_1$. 
It is inherent in BPO that the researcher must introduce non-data based information to distinguish between the two equations.
Given restrictions $\psi(\theta) = 0$, Poirier (1980, JoE) showed [using Rothenberg (1971, Econometrica)] that $\theta$ is locally identified provided

$$\text{rank} \begin{bmatrix} I^{BPO}(\theta) \\ \frac{\partial \psi}{\partial \theta'} \end{bmatrix} = 2K + 1,$$

where the BPO information matrix is
\[ I^{BPO}(\theta) = E_{Z|\theta} \left[ -\frac{\partial^2 L(\theta; z)}{\partial \theta \partial \theta'} \right] \]

\[ = \sum_{i=1}^{n} \frac{1}{p_{Z_n}([1,1]'|\theta)[1-p_{Z_n}([1,1]'|\theta)]} \left[ \frac{\partial P_{Z_n}([1,1]'|\theta)}{\partial \theta} \right] \left[ \frac{\partial P_{Z_n}([1,1]'|\theta)}{\partial \theta'} \right] \]

where

\[ \frac{\partial P_{Z_n}([1,1]'|\theta)}{\partial \theta} = \begin{bmatrix} \frac{\partial P_{Z_n}([1,1]'|\theta)}{\partial \beta_1} \\ \frac{\partial P_{Z_n}([1,1]'|\theta)}{\partial \beta_2} \\ \frac{\partial P_{Z_n}([1,1]'|\theta)}{\partial \rho_{12}} \end{bmatrix} = \begin{bmatrix} \phi(x_n'\beta_1) \Phi \left( \frac{x_n'(\beta_2 - \rho_{12}\beta_1)}{\sqrt{1-\rho_{12}^2}} \right) x_n \\ \phi(x_n'\beta_2) \Phi \left( \frac{x_n'(\beta_1 - \rho_{12}\beta_2)}{\sqrt{1-\rho_{12}^2}} \right) x_n \\ \phi_2(x_n'\beta_1, x_n'\beta_2, \rho_{12}) x_n \end{bmatrix}. \]
• $I^{\text{BPO}}(\theta)$ is a weighted sum across observations of outer products of the gradient vectors $\frac{\partial p_{Z_n}(\omega, \theta)}{\partial \theta}$.

• Provided these gradients exhibit sufficient variations across observations and there is at least one exclusion restriction on $\beta$, $I^{\text{BPO}}(\theta)$ is nonsingular and $\theta$ is locally identified.

  ◦ If the excluded covariate is continuous, then the required variation is obtained.
The restriction $\rho_{12} = 0$ alone may or may not identify $\beta_1$ and $\beta_2$.

The intermediate cases of Meng and Schmidt (1983, IER) (e.g., censored probit) are identified without any restrictions.

- Begin with the simplest case: no covariates.

- Japanese proverb: “Your garden is not complete until there’s nothing more you can take out of it.”
• Pearson (1900, *Phil. Tran. Roy. Soc. of Lon.*, Series A) and Sheppard (1900, *Tran. of the Cam. Phil. Soc.*).

  ○ $K = 1, x_n = 1 \ (n = 1, ..., N)$.

  ○ $\rho_{12}$ is the *tetrachoric correlation coefficient*.

  ○ Now known as *BP with no covariates*.

  ○ With *full observability* everything is identified. Life is great for BP!
Figure 1: BP Likelihood, $\beta_1 = 0$, $\beta_2 = 0$, $\rho = 0$
Figure 2: BP Likelihood Contours, $\beta_1 = 0$, $\beta_2 = 0$, $\rho_{12} = 0$
If \( \rho_{12} \neq 0 \), MLE for BP is more efficient than univariate probit MLEs:

\[
\hat{\beta}_1 = \Phi^{-1}\left(\frac{n_{10} + n_{11}}{N}\right) \quad \hat{\beta}_2 = \Phi^{-1}\left(\frac{n_{01} + n_{11}}{N}\right)
\]

### Cell Counts

<table>
<thead>
<tr>
<th></th>
<th>(y_1 = 0)</th>
<th>(y_1 = 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(y_2 = 0)</td>
<td>(n_{00})</td>
<td>(n_{10})</td>
</tr>
<tr>
<td>(y_2 = 1)</td>
<td>(n_{01})</td>
<td>(n_{11})</td>
</tr>
</tbody>
</table>

The MLE of \( \Phi(Y(y_1 = i, y_2 = j|\theta) \) is \( \hat{p}_{ij} = n_{ij}/N \). The MLE of \( \beta_1 \), \( \beta_2 \), and \( \rho_{12} \) are \( \hat{\beta}_1 = \Phi^{-1}(\hat{p}_{10} + \hat{p}_{11}) \), \( \hat{\beta}_1 = \Phi^{-1}(\hat{p}_{01} + \hat{p}_{11}) \), and (implicitly) \( \hat{p}_{11} = \Phi_2(\hat{\beta}_1, \hat{\beta}_2, \hat{\rho}_{12}) \).
For this intercept-only case, the sample information matrix for BPO is

\[
I_{BPO}^{(\beta_1, \beta_2, \rho_{12})} = \frac{N}{q} \begin{bmatrix}
\phi(\beta_1) \Phi\left(\frac{\beta_2 - \rho_{12} \beta_1}{\sqrt{1 - \rho^2}}\right) & \Phi(\beta_1) \Phi\left(\frac{\beta_2 - \rho_{12} \beta_1}{\sqrt{1 - \rho^2}}\right) \\
\phi(\beta_2) \Phi\left(\frac{\beta_1 - \rho_{12} \beta_2}{\sqrt{1 - \rho^2}}\right) & \Phi(\beta_2) \Phi\left(\frac{\beta_1 - \rho_{12} \beta_2}{\sqrt{1 - \rho^2}}\right) \\
\Phi_2(\beta_1, \beta_2, \rho_{12}) & \Phi_2(\beta_1, \beta_2, \rho_{12})
\end{bmatrix}^{'}
\]

where

\[
q = \Phi_2(\beta_1, \beta_2, \rho_{12}) \left[1 - \Phi_2(\beta_1, \beta_2, \rho_{12})\right].
\]
Clearly \( I^{\text{BPO}}(\theta) \) is singular in this intercept-only case.

- The rank deficiency is 3 - 1 = 2. Knowing \( \rho_{12} \) does not identify \( \beta \).
- \( \beta_1, \beta_2 \) and \( \rho_{12} \) are individually **unidentified** for BPO.
- Only the scalar
  \[
P_{Y_n}([1, 1]'|\theta) = \Phi_2(\beta_1, \beta_2, \rho_{12})
  \]
  is identified. The MLE of \( P_{Y_n}([1, 1]'|\theta) \) is \( \frac{n_{11}}{N} \).
Figure 3: BPO Likelihood, $\beta_1 = 0, \beta_2 = 0, \rho_{12} = 0$
Figure 4: BPO Likelihood Contours, $\beta_1 = 0$, $\beta_2 = 0$, $\rho_{12} = 0$
• Clearly, without restrictions, BPO has a labeling problem.
  ○ In fact, there are hyperplanes through the parameter space for which
    the likelihood is constant.
    - $\beta_2 = \rho_{12} \beta_1$
    - $\beta_1 = \rho_{12} \beta_2$
  ○ Recall:

\[
\frac{\partial P_{Z_n}([1,1]'|\theta)}{\partial \theta} = \begin{bmatrix}
\frac{\partial P_{Z_n}([1,1]'|\theta)}{\partial \beta_1} \\
\frac{\partial P_{Z_n}([1,1]'|\theta)}{\partial \beta_2} \\
\frac{\partial P_{Z_n}([1,1]'|\theta)}{\partial \rho_{12}}
\end{bmatrix}
= \begin{bmatrix}
\phi(x_n'\beta_1) \Phi \left( \frac{x_n' (\beta_2 - \rho_{12} \beta_1)}{\sqrt{1-\rho_{12}^2}} \right) x_n \\
\phi(x_n'\beta_2) \Phi \left( \frac{x_n' (\beta_1 - \rho_{12} \beta_2)}{\sqrt{1-\rho_{12}^2}} \right) x_n \\
\phi_2(x_n'\beta_1, x_n'\beta_2, \rho_{12})
\end{bmatrix}.
\]
A “Peculiar” Case [Poirier (1980, JoE)]

Add a binary regressor to $x_n$ such that $x_n = [1, 0]$ for the first $r$ observations, and $x_n = [1, 1]$ for the last $N - r$ observations.

- Suppose $\beta_1 = [\beta_{11}, 0]'$ and $\beta_2 = [\beta_{21}, \beta_{22}]'$, or in other words,
  \[ \psi(\theta) = \beta_{12} = 0 \text{ and } \frac{\partial \psi}{\partial \theta}' = [0, 1, 0, 0, 0]. \]

- $P_{Z_n}([1, 1]'|\theta) = \Phi_2(\beta_{11}, \beta_{21}, \rho_{12}) \quad (n = 1, \ldots, r)$
- $P_{Z_n}([1, 1]'|\theta) = \Phi_2(\beta_{11}, \beta_{11} + \beta_{12}, \rho_{12}) \quad (n = r + 1, \ldots, N)$. 
• Under $\rho_{12} = 0$ the information matrix for BPO is

$$I^{BPO}(\theta) = \sum_{n=1}^{r} \left( \frac{1}{q_0} \right) \begin{bmatrix} \phi(\beta_{11}) \Phi(\beta_{21}) \\ \phi(\beta_{21}) \Phi(\beta_{11}) \\ \phi(\beta_{21}) \Phi(\beta_{11}) \end{bmatrix} \begin{bmatrix} \phi(\beta_{11}) \Phi(\beta_{21}) \\ \phi(\beta_{21}) \Phi(\beta_{11}) \\ \phi(\beta_{21}) \Phi(\beta_{11}) \end{bmatrix}^\prime +$$

$$\sum_{n=r+1}^{N} \left( \frac{1}{q_1} \right) \begin{bmatrix} \phi(\beta_{11}) \Phi(\beta_{21} + \beta_{22}) \\ \phi(\beta_{21} + \beta_{22}) \Phi(\beta_{11}) \\ \phi(\beta_{21} + \beta_{22}) \Phi(\beta_{11}) \end{bmatrix} \begin{bmatrix} \phi(\beta_{11}) \Phi(\beta_{21} + \beta_{22}) \\ \phi(\beta_{21} + \beta_{22}) \Phi(\beta_{11}) \\ \phi(\beta_{21} + \beta_{22}) \Phi(\beta_{11}) \end{bmatrix}^\prime$$

where

$$q_0 = \Phi(\beta_{11}) \Phi(\beta_{21}) [1 - \Phi(\beta_{11}) \Phi(\beta_{21})],$$

$$q_1 = \Phi(\beta_{11}) \Phi(\beta_{21} + \beta_{22}) [1 - \Phi(\beta_{11}) \Phi(\beta_{21} + \beta_{22})].$$
Define

\[
\delta = \begin{bmatrix}
\frac{\phi(\beta_{21})}{\Phi(\beta_{21})} \\
\frac{\phi(\beta_{11})}{\Phi(\beta_{11})} \\
- \frac{\phi(\beta_{11})}{\Phi(\beta_{11})} \left[ \frac{\phi(\beta_{21}) \Phi(\beta_{21} + \beta_{22})}{\Phi(\beta_{21}) \phi(\beta_{21} \beta_{22})} - 1 \right]
\end{bmatrix}.
\]

- \( \delta' I^{\text{BPO}}(\theta) \delta = 0 \) implying \( I^{\text{BPO}}(\theta) \) is not positive definite.
- Therefore, \( \theta \) is not identified if the excluded covariate is binary.
Recall the “simple” case: there are three parameters $\beta_{11}, \beta_{21}, \rho_{12}$, but only $P_{Z_n}([1,1]′ | \theta)$ is identified with BPO.

- Now add $\beta_{22}$, but there is only one additional cofactor value.
- Even with $\rho_{12} = 0$, $\beta_{11}$, $\beta_{21}$, and $\beta_{22}$ are not individually identified, because the cofactor does not exhibit sufficient variation.
- If the cofactor also took on a third value, then $\beta_{11}$, $\beta_{21}$, and $\beta_{22}$ are individually identified if $\rho_{12} = 0$.
- If the cofactor is continuous, then no peculiar identification problem arises.
- The non-linearity of $\mathcal{L}^{BPO}(\theta)$ provides the local identification.
\[ \frac{\partial L^{\text{TPO}}(\theta; z)}{\partial \theta} = \sum_{n=1}^{N} \frac{z_n - \Phi_3(\mu; \rho)}{\left[1 - \Phi_3(\mu; \rho)\right] \Phi_3(\mu; \rho)} \left[ \frac{\partial \Phi_3(\mu; \rho)}{\partial \theta} \right], \]

where

\[ \frac{\partial \Phi_3(\mu_n; \rho)}{\partial \theta} = \begin{bmatrix}
\frac{\partial \Phi_3(\mu_n; \rho)}{\partial \beta_1} \\
\frac{\partial \Phi_3(\mu_n; \rho)}{\partial \beta_2} \\
\frac{\partial \Phi_3(\mu_n; \rho)}{\partial \beta_3} \\
\frac{\partial \Phi_3(\mu_n; \rho)}{\partial \rho_{12}} \\
\frac{\partial \Phi_3(\mu_n; \rho)}{\partial \rho_{13}} \\
\frac{\partial \Phi_3(\mu_n; \rho)}{\partial \rho_{23}}
\end{bmatrix} = \begin{bmatrix}
\Phi(\mu_{n1}) \Phi_2(v_{n2,1}, v_{n3,1}; \rho_{23,1}) x_n \\
\Phi(\mu_{n2}) \Phi_2(v_{n1,2}, v_{n3,2}, \rho_{13,2}) x_n \\
\Phi(\mu_{n3}) \Phi_2(v_{n1,3}, v_{n2,3}, \rho_{12,3}) x_n \\
\Phi_2(\mu_{n1}, \mu_{n2}; \rho_{12}) \Phi \left( \frac{(1 - \rho_{12}^2) \mu_{n3} - (\rho_{13} - \rho_{12} \rho_{23}) \mu_{n1} - (\rho_{23} - \rho_{12} \rho_{13}) \mu_{n2}}{\sqrt{|\Omega(\rho)| \cdot (1 - \rho_{12}^2)}} \right) \\
\Phi_2(\mu_{n2}, \mu_{n3}; \rho_{23}) \Phi \left( \frac{(1 - \rho_{23}^2) \mu_{n2} - (\rho_{12} - \rho_{13} \rho_{23}) \mu_{n1} - (\rho_{23} - \rho_{13} \rho_{12}) \mu_{n3}}{\sqrt{|\Omega(\rho)| \cdot (1 - \rho_{23}^2)}} \right) \\
\Phi_2(\mu_{n1}, \mu_{n3}; \rho_{13}) \Phi \left( \frac{(1 - \rho_{13}^2) \mu_{n1} - (\rho_{12} - \rho_{23} \rho_{13}) \mu_{n2} - (\rho_{13} - \rho_{23} \rho_{12}) \mu_{n3}}{\sqrt{|\Omega(\rho)| \cdot (1 - \rho_{13}^2)}} \right)
\end{bmatrix}. \]
Identification depends on the nonsingularity of

\[ I^{MPO}(\theta) = \sum_{n=1}^{N} \frac{1}{\Phi_3(\mu_n; \rho) [1 - \Phi_2(\mu_n; \rho)]} \left[ \frac{\partial \Phi_3(\mu_n; \rho)}{\partial \theta} \right] \left[ \frac{\partial \Phi_3(\mu_n; \theta)}{\partial \theta'} \right], \tag{17} \]

which rests on sufficient variation of the gradient over observations.

- In the intercept-only case the only thing identified is the scalar \( \Phi_3(\beta_1, \beta_2, \beta_3; \rho_{12}, \rho_{23}, \rho_{13}) \).

- Unlike the rank deficiency of two in the bivariate case, the rank deficiency is \( 6 - 1 = 5 \) in the trivariate case.
Also there are labeling problems with information matrix (17) becoming singular if \( \beta_2 = \rho_{12} \beta_1, \ \beta_1 = \rho_{12} \beta_2, \ \beta_3 = \rho_{13} \beta_1, \ \beta_1 = \rho_{13} \beta_3, \ \beta_2 = \rho_{23} \beta_3, \) and \( \beta_3 = \rho_{23} \beta_2. \)

Following the discussion of the BPO case, let \( x_n = [1, x_{n2}, x_{n3}]' \) and suppose \( \beta_1 = [\beta_{11}, 0, 0]', \ \beta_2 = [\beta_{21}, \beta_{22}, 0]', \) and \( \beta_3 = [\beta_{31}, 0, \beta_{33}]'. \)

- In other words, two continuous covariates \( x_{n2} \) and \( x_{n3} \) are introduced together with the four restrictions \( \beta_{12} = \beta_{13} = \beta_{23} = \beta_{32} = 0. \)

- For the moment also suppose \( \rho = 0_3. \)
Dropping the last three elements in (36) yields

\[
\begin{bmatrix}
\frac{\partial \Phi_3(\mu_n; \rho)}{\partial \beta_{11}} \\
\frac{\partial \Phi_3(\mu_n; \rho)}{\partial \beta_{21}} \\
\frac{\partial \Phi_3(\mu_n; \rho)}{\partial \beta_{22}} \\
\frac{\partial \Phi_3(\mu_n; \rho)}{\partial \beta_{31}} \\
\frac{\partial \Phi_3(\mu_n; \rho)}{\partial \beta_{33}}
\end{bmatrix}
= \begin{bmatrix}
\Phi(\beta_{11}) \Phi(\beta_{21} + \beta_{22}x_{n2}) \Phi(\beta_{31} + \beta_{33}x_{n3}) \\
\Phi(\beta_{11}) \Phi(\beta_{21} + \beta_{22}x_{n2}) \Phi(\beta_{31} + \beta_{33}x_{n3}) \\
\Phi(\beta_{11}) \Phi(\beta_{21} + \beta_{22}x_{n2}) \Phi(\beta_{31} + \beta_{33}x_{n3})x_{n2} \\
\Phi(\beta_{11}) \Phi(\beta_{31} + \beta_{33}x_{n3}) \Phi(\beta_{21} + \beta_{22}x_{n2}) \\
\Phi(\beta_{11}) \Phi(\beta_{31} + \beta_{33}x_{n3}) \Phi(\beta_{21} + \beta_{22}x_{n2})x_{n3}
\end{bmatrix}.
\]
Computation Issues

• The MP likelihood function requires computation of $\Phi_j(\mu_n; \rho)$.
• $J \leq 4$ in applications in health, labor or education economics.
• Lesaffre and Kaufman (1992, *JASA*) showed
  □ $\Phi_j(\mu_n; \rho)$ is strictly concave in $\beta$ given $\rho$.
  □ $\Phi_j(\mu_n; \rho)$ is not strictly concave in $\rho$.
  □ MLE is consistent asymptotically normal provided no perfect classifiers exist.
  □ Did not really address computational issues.

\[ \Phi_2(0, 0; \rho_{12}) = \frac{1}{2} + \frac{\arccos(\rho_{12})}{2\pi} \]

- strictly monotone in \( \rho_{12} \).
- convex on (-1, 0)
- concave on (0, 1)

\[ \Phi_3(0, 0; \rho_{12}, \rho_{23}, \rho_{13}) = \frac{1}{2} + \frac{\arccos(\rho_{12}) + \arccos(\rho_{23}) + \arccos(\rho_{13})}{4\pi} \]
Huguenin, Pelgrin and Holly (2009) provide an exact decomposition of the cdf of the J-variate (standardized) normal vector encountered in the likelihood function of a multivariate probit model.

- Obtain a sum of multivariate integrals, in which the highest dimension of the integrands is $J - 1$.
  - Integration domains are bounded - simplifying the integration.
  - Based on Plackett (1954, *Biometrika*).
- $J = 4$ is feasible.
- Also consider the singular case.
• Gassmann (2003, *JCGS*) also employed a recursive numerical method based on Plackett (1954, *Biometrika*).
  
  ◦ Argues it works for $J \leq 10$.
  

- Computation times depend on $J$, the correlation structure, the magnitude of the sought probability, and the required accuracy.

- Numerical tests were conducted on approximately 3,000 problems generated randomly in up to $J = 20$ dimensions.
The findings indicate that direct integration methods give acceptable results for up to $J = 12$, provided that the probability mass of the rectangle is not too large (less than about 0.9).

For problems with small probabilities (less than 0.3) a crude Monte Carlo method gives reasonable results quickly, while bounding procedures perform best on problems with large probabilities ($> 0.9$).

For larger problems numerical integration with quasi-random Korobov points and a decomposition method due to Deak.
• Factor structures for the correlation matrix.
  ◦ Integration problems are typically greatly reduced.
  ◦ Muthén (1979, *JASA*): $J = 4$
• Bayesians must sample from non-tractable posterior that is proportional to the product of the prior and the MP likelihood.


• It is possible to sample from the posterior without computing the likelihood function!
Bayesian examples:

- Duvvuri and Gruca (2010, *Psychometrika*): $J = 4$
- Hahn, Carvallo and Scott (2010): $J = 100$, six factors
• MCMC methods have difficulty dealing with the generation of draws from the correlation matrix, because its conditional density is of nonstandard form.

○ Methods have been developed to generate draws using variants of

○ These approaches are computationally intensive and therefore are of limited use when applied to large dimensional problems.
• One contested issue is whether to work in terms of
  
  
  ◦ identified correlation matrix [Chib and Greenberg (1998, Biometrika)]
Priors

- Given the information loss in BPO, an informative Bayesian analysis is an attractive option for supplementing sample information.
  - Identification of $\theta$ in the BPO model is often weak, and what is needed is an informative prior with some public appeal, not a “noninformative” prior.
  - Below is a prior family which should be attractive in many situations.
  - Later I will suggest a restricted version depending on only four easily interpreted hyperparameters.
• Assume $\beta = [\beta_1', \beta_2']' \perp \rho$ with joint pdf

$$f(\beta_1, \beta_2, \rho \mid b, V, a) = \Phi_{K_1+K_2}(\beta \mid b, V) \, f_{\beta}(\rho \mid a).$$

where $K = K_1 + K_2$.

○ This is a conjugate multivariate normal prior for $\beta$ given $\rho$.

○ The prior pdf $f_{\beta}(\rho \mid a)$ is the symmetric beta

$$f_{\beta}(\rho \mid a) = \frac{\Gamma(2a)}{[\Gamma(a)]^2} \rho^{a-1}(1 - \rho)^{a-1}, \quad -1 < \rho < 1.$$
Posterior

- The posterior pdf of $\theta$ is unfortunately analytically intractable. So Bayesian analysis rests on the ability to draw a sample from it.

  - BPO can be viewed as a binary outcome model in which the link function is a bivariate standard normal cdf.

  - Thus consider the \textit{data augmented posterior} $\beta, \rho, z^* | z$, where $z^* = [y_{11}^*, ..., y_{N_1}^*, y_{12}^*, ..., y_{N_2}^*]'$. 
Set $\beta^{(0)} = 0_K$, $\rho^{(0)} = 0$, and $r = 0$ and proceed as follows.

**Step 1:** Consider the distribution $z^* \mid z, \beta, \rho$. The only difference from the standard data augmentation case is that we must sample from a truncated *bivariate* distribution here. Let

$$
\Sigma = \Sigma(\rho) = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.
$$
Set $r = r + 1$ and sample either

$$z^{(r)} | z, \beta^{(r-1)}, \rho^{(r-1)} \sim \mathbb{N}_{2N} \left( \begin{bmatrix} X_1 \beta_1^{(r-1)} \\ X_2 \beta_2^{(r-1)} \end{bmatrix}, \Sigma (\rho^{(r-1)}) \otimes I_2 \right)$$

truncated to quadrants other than the upper right if $z_n = 0$.

or

$$z^{(r)} | z, \beta^{(r-1)}, \rho^{(r-1)} \sim \mathbb{N}_{2N} \left( \begin{bmatrix} X_1 \beta_1^{(r-1)} \\ X_2 \beta_2^{(r-1)} \end{bmatrix}, \Sigma (\rho^{(r-1)}) \otimes I_2 \right)$$

truncated to the upper right quadrant if $z_n = 1$, 

85
Step 2: \( \beta \mid z^*, z, \rho \sim \mathbb{N}_K(\mathbf{b}^{(r)}, \mathbf{V}^{(r)}) \), where

\[
[\mathbf{V}^{-1}]^{(r)} = \mathbf{V}^{-1} + \begin{bmatrix} X_1' & 0 \\ 0 & X_2' \end{bmatrix} [\Sigma (\rho^{(r-1)}) \otimes I_2]^{-1} \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix},
\]

\[
\mathbf{b}^{(r)} = [\mathbf{V}^{-1}]^{(r)} \left( \mathbf{V}^{-1} \mathbf{b} + \begin{bmatrix} X_1' & 0 \\ 0 & X_2' \end{bmatrix} [\Sigma (\rho^{(r-1)}) \otimes I_2]^{-1} \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \beta^{*(r)} \right),
\]

\[
\beta^{*(r)} = \left( \begin{bmatrix} X_1' & 0 \\ 0 & X_2' \end{bmatrix} [\Sigma (\rho^{(r-1)}) \otimes I_2]^{-1} \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \right)^{-1} \begin{bmatrix} X_1' & 0 \\ 0 & X_2' \end{bmatrix} [\Sigma (\rho^{(r-1)}) \otimes I_2]^{-1} z^{*(r)}.
\]
Step 3: $\rho \mid z^*, z, \beta$, where

$$f(\rho \mid z^*, z, \beta) \propto f(\rho) \phi_{2N} \left( z^* \left| \begin{bmatrix} X_1 \beta_1 \\ X_2 \beta_2 \end{bmatrix}, \Sigma(\rho) \otimes I_N \right. \right), \quad -1 < \rho < 1.$$ 

The random walk Metropolis-Hastings algorithm is a convenient way to sample from this univariate distribution.
Write $\rho^{\dagger(r)} = \rho^{(r - 1)} + u$, where $\rho^{\dagger(r)}$ is the candidate value, $\rho^{(r - 1)}$ is the current value, and $u \sim \mathcal{N}(0, N^{-1})$.

Let

$$\delta = \begin{bmatrix}
\frac{f_{\beta}(\rho^{\dagger(r)} | \mathbf{a})}{f_{\beta}(\rho^{(r - 1)} | \mathbf{a})} \\
\frac{\phi_{2N}(z^* | [X_1 \beta_1^{(r)}'), (X_2 \beta_2^{(r)})', \Sigma(\rho^{\dagger(r)} \otimes I_N)]}{\phi_{2N}(z^* | [X_1 \beta_1^{(r - 1)}'), (X_2 \beta_2^{(r - 1)})', \Sigma(\rho^{(r - 1)} \otimes I_N)]}
\end{bmatrix}.$$ 

Generate a proposed value $\rho^{\dagger(r)}$ by drawing $u$. Put $\rho^{(r)} = \rho^{\dagger(r)}$ with probability

$$\alpha(\rho^{\dagger}, \rho) = \min \{ \delta, 1 \}$$

and leave $\rho^{(r)} = \rho^{(r - 1)}$ with probability $1 - \alpha(\rho^{\dagger}, \rho)$. 
• Go back to step 1 and repeat until convergence is obtained. As $R \to \infty$, the resulting draws converge in distribution to the augmented posterior.

• Given a sample $z^{(r)}, \beta^{(r)}, \rho^{(r)} (r = 1, 2, ..., R)$, it is then possible to compute posterior expectations of a quantity of interest, say, $h(\beta_1, \beta_2, \rho)$. Examples are:
  ◦ Moments of parameters.
  ◦ Posterior probabilities of the signs of parameters.
  ◦ Predictive density $f(\tilde{y} | \tilde{x}_1, \tilde{x}_1, \tilde{x}_2, z)$ corresponding to covariates $\tilde{x}_1$ and $\tilde{x}_2$, and $h(\beta_1, \beta_2, \rho) = \tilde{p}^{ij}(\tilde{x}_1, \tilde{x}_2) = \text{Prob} (\tilde{y}_1 = i \text{ and } \tilde{y}_2 = j)$ ($i, j = 0, 1$).
Restricted Prior

• The following restrictions should be attractive in many cases. Their appeal rests on the easiness of choosing the hyperparameters.
• Suppose

  ◦ the first element in $x_{n1}$ and $x_{n2}$ is unity,

  ◦ any continuous covariates are measured as deviations from their sample means divided by sample standard deviations, and

  ◦ dummy variables are left as is.

  ◦ These *conventions* guarantee a simple interpretation for the intercepts and render coefficients unitless facilitating prior elicitation.
- Suppose \( \mathbf{b} = [\mathbf{b}_1 \mathbf{e}_1', \mathbf{b}_2 \mathbf{e}_2']' \), where \( \mathbf{e}_i = [1, \mathbf{0}_{K_i-1}']' \) (\( i = 1, 2 \)), \( \mathbf{c} > 0 \), and

\[
\mathbf{V}^{-1} = \begin{bmatrix}
\mathbf{V}_1 & \mathbf{0}_{K_1 \times K_2} \\
\mathbf{0}_{K_2 \times K_1} & \mathbf{V}_2
\end{bmatrix} = \mathbf{c} \begin{bmatrix}
\mathbf{X}_1' \mathbf{X}_1 & \mathbf{0}_{K_1 \times K_2} \\
\mathbf{0}_{K_2 \times K_1} & \mathbf{X}_2' \mathbf{X}_2
\end{bmatrix}.
\]

- This prior centers the slopes over zero, leaving specification of the hyperparameters \( \mathbf{b}_1 \) and \( \mathbf{b}_2 \) to center the intercepts.
- One way of choosing \( \mathbf{b}_j \) (\( j = 1, 2 \)) is to consider the average observation \( \bar{x}_j = [1, 0, ..., 0]' \), subjectively assess \( \mathbf{p}_j = \text{Prob}(y_{nj} = 1) \), and then choose \( \mathbf{b}_j = \Phi^{-1}(\mathbf{p}_j) \) (\( j = 1, 2 \)).
- The hyperparameter \( \mathbf{c} \) controls the prior precision for \( \beta \) relative to the precision in OLS applied separately to the latent system.
• The remaining two hyperparameters \( a \) and \( c \) are candidates for a sensitivity analysis reflecting the strength of the prior.

  ° The hyperparameter \( a \) determines the precision \( 4(2a + 1) \) of \( \rho \).

  ° Centering the prior for \( \rho \) over zero (implying the two binary decisions are independent), is often a convenient reference point.

  ° This prior has many similarities with priors in Adkins and Hill (1996), Zellner (1986), and Zellner and Rossi (1984, \textit{JoE}), but applied here to \textit{bivariate} probit.
Note $\beta_j | y_j^* \sim \mathcal{N}(\bar{b}_j, \bar{V}_j)(j = 1, 2)$, where
\[
\bar{V}_j^{-1} = (\bar{V}_j^{-1} + X_j'X_j) = (c + 1)X_j'X_j,
\]
\[
\bar{b}_j = \bar{V}_j^{-1}(\bar{V}_j^{-1}b_1 + X_j'X_jb_j) = (cb_j + b_j),
\]
\[
y_j^* = [y_{1j}^*, y_{2j}^*, ..., y_{Nj}^*]'
\]
\[
b_j = (X_j'X_j)^{-1}X_jy_j^*.
\]

- Posterior mean slopes shrink the OLS slopes toward $0_{K_j-1}$ and the posterior mean intercept shrinks the OLS slope toward $b_j$.

- Centering the slopes over non-zero values, or changing prior precision matrix to an arbitrary $V_j^{-1}(j = 1, 2)$ does not complicate matters.
Concluding Comments

- multivariate-t [Chib (2000), Dey and Chen (1996, 2008)]

- Chen and Dey (1998, *Sankhyā*): scale mixture of multivariate normal link.

- Chen and Shao (1999, *JMA*) considered multivariate data where some are binary and others are ordinal; Bayesian, $J = 3$.

- multivariate multinomial [Zhang, Boscardin, and Belin (2008, *CSDA*)]
Correlated binary response data with covariates:

- Liang and Zeger (1986, *Biometrika*)
- Zeger and Liang (1986, *Biometrics*)
- Prentice (1988, *Biometrics*)
- Tan, Qu and Kutner (1997, *Comm. Stat.*)
- Carey, Zeger and Diggle (1993, *Biometrika*)