Identification in Multinomial Probit Models with Associated Outcomes and Measurements

James J. Heckman

Revised, November 4, 2011
Index System:

\[ I_1 = \phi_1 W_1 + X \beta_1 + U_1 \]
\[ I_2 = \phi_2 W_2 + X \beta_2 + U_2 \]
\[ I_3 = \phi_3 W_3 + X \beta_3 + U_3 \]  

Wage System:

\[ W_1 = Z_1 \gamma_1 + \varepsilon_1 \]
\[ W_2 = Z_2 \gamma_2 + \varepsilon_2 \]
\[ W_3 = Z_3 \gamma_3 + \varepsilon_3 \]
Substitute wages into index equations

\begin{align*}
I_1 &= Z_1(\phi_1 \gamma_1) + X \beta_1 + (U_1 + \phi_1 \epsilon_1) \\
I_2 &= Z_2(\phi_2 \gamma_2) + X \beta_2 + (U_2 + \phi_2 \epsilon_2) \\
I_3 &= Z_3(\phi_3 \gamma_3) + X \beta_3 + (U_3 + \phi_3 \epsilon_3)
\end{align*}

(3)

We need to preserve cross equation restrictions.
Measurements of factor structure:

\[ M_1 = \mu_1 + \lambda_1^m \theta + V_1^M \]
\[ M_2 = \mu_2 + \lambda_2^m \theta + V_2^M \]
\[ \vdots \quad \vdots \]
\[ M_J = \mu_J + \lambda_J^m \theta + V_J^M \]

\[ M = \mu + \Lambda^M \theta + V^M \]
\[ \varepsilon = \Lambda^\varepsilon \theta + V^\varepsilon \]
\[ U = \Lambda^U \theta + V^U \]

Components of \( V^\varepsilon \), \( V^U \), \( V^M \) are mutually uncorrelated within each vector and across each vector and with \( \theta \).
Choice System:

The agent chooses the $j$ that maximizes the index:

$$\hat{j} = \arg \max \{ I_j \}_{j=1}^3.$$

Consider (1), (2), (3) and ignore the factor structure for the moment.

(a) Suppose we observe the $W_j$ and that they are exogenous and that $W_j$ can be freely varied against $W_{j'}$. 

James J. Heckman

ID in Multinomial Probit Models
For $\hat{j} = 1$,

\[ I_1 > I_2 \iff \phi_1 W_1 - \phi_2 W_2 + X(\beta_1 - \beta_2) + U_1 - U_2 > 0 \]
\[ I_1 > I_3 \iff \phi_1 W_1 - \phi_3 W_3 + X(\beta_1 - \beta_3) + U_1 - U_3 > 0 \]

For $\hat{j} = 2$,

\[ I_2 > I_1 \iff \phi_2 W_2 - \phi_1 W_1 + X(\beta_2 - \beta_1) + U_2 - U_1 > 0 \]
\[ I_2 > I_3 \iff \phi_2 W_2 - \phi_3 W_3 + X(\beta_2 - \beta_3) + U_2 - U_3 > 0 \]

For $\hat{j} = 3$,

\[ I_3 > I_1 \iff \phi_3 W_3 - \phi_1 W_1 + X(\beta_3 - \beta_1) + U_3 - U_1 > 0 \]
\[ I_3 > I_2 \iff \phi_3 W_3 - \phi_2 W_2 + X(\beta_3 - \beta_2) + U_3 - U_2 > 0 \]

\[ U_1 - U_2 = -(U_2 - U_1) \]
\[ U_1 - U_3 = -(U_3 - U_1) \]

etc.
From Bivariate Probits applied to each relationship, we can form the left hand side of the following relationship:

\[
\Pr \left[ j = 1 \mid . \right] = \int_{-\infty}^{\phi_1 W_1 - \phi_2 W_2 + X(\beta_1 - \beta_2)} \frac{1}{\sqrt{2\pi \sigma_a \sigma_b (1 - \rho_1^2)}} \exp \left( \frac{a^2 + b^2 - 2\rho_1 ab}{-2(1 - \rho_1^2)} \right) dadb,
\]

where \( a = \frac{U_2 - U_1}{\text{var}(U_1 - U_2)^{\frac{1}{2}}} \) and \( b = \frac{U_3 - U_1}{\text{var}(U_1 - U_3)^{\frac{1}{2}}} \).

\( \sigma_a = \sqrt{\text{Var}(a)} \) \( \sigma_b = \sqrt{\text{Var}(b)} \).
Symmetrically We Have Other Expressions

Consider Evidence From The First Choice

We identify

\[
\begin{align*}
\frac{\beta_1 - \beta_2}{\sqrt{\text{var}(U_1 - U_2)}} & \quad \frac{\beta_1 - \beta_3}{\sqrt{\text{var}(U_1 - U_3)}} & \quad \frac{\phi_1}{\sqrt{\text{var}(U_1 - U_2)}} \\
\frac{\phi_1}{\sqrt{\text{var}(U_1 - U_3)}} & \quad \frac{\phi_2}{\sqrt{\text{var}(U_1 - U_2)}} & \quad \frac{\phi_3}{\sqrt{\text{var}(U_1 - U_3)}}
\end{align*}
\]

as well as

\[
\text{correl}(U_1 - U_2, U_1 - U_3)
\]

(See Plackett, 1954, as cited in the attached notes for identification of the correlation coefficient)

If \( \phi_1 \neq 0 \), we identify \( \frac{\text{var}(U_1 - U_3)}{\text{var}(U_1 - U_2)} \) from the ratio of the two normalized \( \phi_j \).
Second Choice

\[ \frac{\beta_2 - \beta_1}{\sqrt{\text{var}(U_2 - U_1)}} \text{,} \quad \frac{\beta_2 - \beta_3}{\sqrt{\text{var}(U_2 - U_3)}} \text{,} \quad \frac{\phi_2}{\sqrt{\text{var}(U_1 - U_2)}} \text{,} \quad \frac{\phi_3}{\sqrt{\text{var}(U_2 - U_3)}} \text{,} \]

\[ \text{correl}(U_2 - U_1, U_1 - U_3) \]

\[ \phi_2 \neq 0 \implies \text{we can identify} \frac{\text{var}(U_1 - U_3)}{\text{var}(U_2 - U_3)} \]
Third Choice:
Determined from other two.

\[
\begin{align*}
\frac{\beta_3 - \beta_1}{\sqrt{\text{var}(U_3 - U_1)}} & , \\
\frac{\beta_3 - \beta_2}{\sqrt{\text{var}(U_3 - U_2)}} & , \\
\frac{\phi_3}{\sqrt{\text{var}(U_3 - U_2)}} & , \\
\frac{\phi_1}{\sqrt{\text{var}(U_3 - U_1)}} & , \\
\frac{\phi_2}{\sqrt{\text{var}(U_3 - U_2)}} & , \\
\phi_3 & \neq 0 \implies \text{we can identify} \frac{\text{var}(U_1 - U_3)}{\text{var}(U_3 - U_2)}
\end{align*}
\]
I. Suppose $U_1, U_2, U_3$ are independent

For $\hat{j} = 1$, we identify

$$\frac{\beta_1 - \beta_2}{(\sigma_1^2 + \sigma_2^2)^{1/2}}, \quad \frac{\beta_1 - \beta_3}{(\sigma_1^2 + \sigma_3^2)^{1/2}}, \quad \frac{\phi_1}{(\sigma_1^2 + \sigma_2^2)^{1/2}},$$

$$\frac{\phi_1}{(\sigma_1^2 + \sigma_3^2)^{1/2}}, \quad \frac{\phi_2}{(\sigma_1^2 + \sigma_2^2)^{1/2}}, \quad \frac{\phi_3}{(\sigma_1^2 + \sigma_3^2)^{1/2}}$$

$$\rho_1 = \text{correl}(U_1 - U_2, U_1 - U_3) = \frac{\sigma_1^2}{(\sigma_1^2 + \sigma_2^2)^{1/2}(\sigma_1^2 + \sigma_3^2)^{1/2}}$$

$$\frac{\sigma_1^2 + \sigma_3^2}{\sigma_1^2 + \sigma_2^2} \quad \text{identified from the ratio of the two normalized } \phi_1 \quad \text{coefficients.}$$
Since I know the ratio, I can use first two terms to identify

\[
\frac{\beta_2 - \beta_3}{(\sigma_1^2 + \sigma_2^2)^{\frac{1}{2}}}, \quad \frac{\beta_1 - \beta_3}{(\sigma_1^2 + \sigma_3^2)^{\frac{1}{2}}}
\]

For \( \hat{j} = 2 \), we identify

\[
\rho_2 = \frac{\sigma_2^2}{(\sigma_1^2 + \sigma_2^2)^{\frac{1}{2}}(\sigma_2^2 + \sigma_3^2)^{\frac{1}{2}}}
\]
\[
\begin{align*}
\frac{\beta_2 - \beta_1}{(\sigma_1^2 + \sigma_2^2)^{\frac{1}{2}}} & , \quad \frac{\beta_2 - \beta_3}{(\sigma_2^2 + \sigma_3^2)^{\frac{1}{2}}} , \quad \frac{\phi_2}{(\sigma_1^2 + \sigma_2^2)^{\frac{1}{2}}} , \\
& , \quad \frac{\phi_2}{(\sigma_2^2 + \sigma_3^2)^{\frac{1}{2}}} , \quad \frac{\phi_1}{(\sigma_1^2 + \sigma_2^2)^{\frac{1}{2}}} , \quad \frac{\phi_3}{(\sigma_2^2 + \sigma_3^2)^{\frac{1}{2}}}
\end{align*}
\]

Knowledge of \( \frac{\sigma_1^2 + \sigma_3^2}{\sigma_1^2 + \sigma_2^2} \) \( \Rightarrow \) I know \( \frac{\beta_1 - \beta_3}{(\sigma_2^2 + \sigma_3^2)^{\frac{1}{2}}} \) and \( \frac{\beta_1 - \beta_3}{(\sigma_1^2 + \sigma_3^2)^{\frac{1}{2}}} \)
For $j = 3$, we identify

$$
\rho_3 = \frac{\sigma_3^2}{(\sigma_2^2 + \sigma_3^2)^{1/2}(\sigma_3^2 + \sigma_1^2)^{1/2}}
$$

$$
\frac{\beta_3 - \beta_1}{(\sigma_3^2 + \sigma_1^2)^{1/2}}, \quad \frac{\beta_3 - \beta_2}{(\sigma_3^2 + \sigma_2^2)^{1/2}}, \quad \frac{\phi_3}{(\sigma_1^2 + \sigma_3^2)^{1/2}}
$$

$$
\frac{\phi_3}{(\sigma_3^2 + \sigma_2^2)^{1/2}}, \quad \frac{\phi_1}{(\sigma_1^2 + \sigma_2^2)^{1/2}}, \quad \frac{\phi_2}{(\sigma_2^2 + \sigma_3^2)^{1/2}}, \quad \frac{\sigma_3^2 + \sigma_2^2}{\sigma_1^2 + \sigma_2^2}
$$
Suppose that I set $\sigma_1^2 = 1$, then from $\rho_1$ we know

$$
\frac{1}{(1 + \sigma_2^2)^\frac{1}{2}(1 + \sigma_3^2)^\frac{1}{2}}
$$

and

$$
\frac{(1 + \sigma_3^2)^\frac{1}{2}}{(1 + \sigma_2^2)^\frac{1}{2}}
$$

∴ know $\sigma_2^2$ and $\sigma_3^2$

∴ we know $\beta_1 - \beta_2$, $\beta_1 - \beta_3$

Set $\beta_3 = 0$ (this is necessary). All of this is normalized with respect to $\text{Var}(U_1) = 1$. 
An Income-Maximizing Model

Notice also that under an income-maximizing model
\[ \phi_1 = \phi_2 = \phi_3 = 1 \] (a la Rosen and Willis), the model is identified. The scale is set by setting the value units — all changes in units are known. We can work with other units, but the key points is that we know the units, i.e. we assume \( \phi_1 = \phi_2 = \phi_3 \) is known. (Two parts: equality and parameter known)

(Why ?) — we get \( (\sigma_1^2 + \sigma_3^2) \), \( (\sigma_1^2 + \sigma_2^2) \) from first choice.

Symmetrically \( (\sigma_2^2 + \sigma_3^2) \) from the second choice.

\[ \therefore \text{We get } \sigma_1^2 \]

\[ \therefore \text{We can identify all of the variances} \]

We do not need to normalize \( \text{var}(U_1) = 1 \).

We still need a normalization on \( \beta_j \) e.g. \( \beta_3 = 0 \).
One Factor Case:

Take the one factor case for simplicity

\[ U_1 = \tau_1 \theta + \omega_1 \]
\[ U_2 = \tau_2 \theta + \omega_2 \]
\[ U_3 = \tau_3 \theta + \omega_3 \]

\( \omega_i \) mutually independent and independent of \( \theta \)

Go back to the case I but add \( \theta \) (condition on it). We can form Bartlett scores.
When we condition on $\theta$, we estimate

$$
\frac{\tau_1 - \tau_2}{(\sigma^2_{\omega_1} + \sigma^2_{\omega_2})^{\frac{1}{2}}} \quad \frac{\tau_1 - \tau_3}{(\sigma^2_{\omega_1} + \sigma^2_{\omega_3})^{\frac{1}{2}}} \quad \frac{\tau_2 - \tau_3}{(\sigma^2_{\omega_2} + \sigma^2_{\omega_3})^{\frac{1}{2}}}
$$

Otherwise it is as before (everything else goes through as before). Replacing $U_i$ by $\omega_i$ (this assumes we know the $\theta$). All previous analysis applies. Thus, in the income-maximizing case, we can identify $\sigma^2_{\omega_j}$.
Can we identify $\sigma^2_{\omega_j}$ in the general case where we condition on $\theta$?

We have 2 (nonlinear) equations in 2 unknowns. ($\rho_3$ is determined given $\rho_1$ and $\rho_2$.)

\[
\begin{align*}
\rho_1(\sigma^2_{\omega_1}, \sigma^2_{\omega_2}) \\
\rho_2(\sigma^2_{\omega_1}, \sigma^2_{\omega_2})
\end{align*}
\]

Need to check solutions.

Compute Jacobian

\[
\begin{pmatrix}
\frac{\partial \rho_1}{\partial \sigma^2_{\omega_1}} & \frac{\partial \rho_1}{\partial \sigma^2_{\omega_2}} \\
\frac{\partial \rho_2}{\partial \sigma^2_{\omega_1}} & \frac{\partial \rho_2}{\partial \sigma^2_{\omega_2}}
\end{pmatrix}
\]

Check if nonsingular.
Set $\tau_3 = 0$, $\sigma_{\omega_3} = 1$

We get

\[
\begin{align*}
\tau_2 &= \frac{\tau_2}{(\sigma^2_{\omega_2} + 1)^{\frac{1}{2}}} \\
\tau_1 &= \frac{\tau_1}{(\sigma^2_{\omega_1} + 1)^{\frac{1}{2}}} \\
\tau_1 - \tau_2 &= \frac{\tau_1 - \tau_2}{(\sigma^2_{\omega_1} + \sigma^2_{\omega_2})^{\frac{1}{2}}}
\end{align*}
\]

From the bivariate probit correlation index we can recover

\[
\rho_j = \frac{\sigma^2_{\omega_j}}{(\sigma^2_{\omega_j} + \sigma^2_{\omega_2})^{\frac{1}{2}} (\sigma^2_{\omega_2} + 1)^{\frac{1}{2}}}
\]

where $\rho_j$ is the correlation associated with the $j^{th}$ choice. We need to solve the system: Check when rank condition is satisfied.
Not conditioning on $\theta_j$, we get

\[ U_1 - U_3 = (\tau_1 - \tau_3)\theta + (\varepsilon_1 - \varepsilon_3) \]
\[ U_1 - U_2 = (\tau_1 - \tau_2)\theta + (\varepsilon_1 - \varepsilon_2) \]

\[
\text{correl}(U_1 - U_3, U_1 - U_2) = \frac{(\tau_1 - \tau_3)(\tau_1 - \tau_2)\sigma^2_\theta + \sigma^2_{\varepsilon_1}}{\left[ (\tau_1 - \tau_3)^2 \sigma^2_\theta + \sigma^2_{\varepsilon_1} + \sigma^2_{\varepsilon_3} \right]^{1/2}} \left[ (\tau_1 - \tau_2)^2 \sigma^2_\theta + \sigma^2_{\varepsilon_1} + \sigma^2_{\varepsilon_3} \right]^{1/2}
\]

\[
\text{correl}(U_2 - U_3, U_2 - U_1) = \frac{(\tau_2 - \tau_3)(\tau_2 - \tau_1)\sigma^2_\theta + \sigma^2_{\varepsilon_2}}{\left[ (\tau_2 - \tau_3)^2 \sigma^2_\theta + \sigma^2_{\varepsilon_2} + \sigma^2_{\varepsilon_3} \right]^{1/2}} \left[ (\tau_2 - \tau_1)^2 \sigma^2_\theta + \sigma^2_{\varepsilon_2} + \sigma^2_{\varepsilon_1} \right]^{1/2}
\]
Additional Notes

In place of the exogeneity assumption on $W_j$, assume it on $Z_j$ and assume at least one exclusion in each equation.

From selection equation, we can estimate (using standard selection equation)

$$\text{var}(\varepsilon_i) \quad i = 1, \ldots, 3, \quad V_i \quad i = 1, \ldots, 3$$

Issues to settle –
Can we identify $\text{cov}(\varepsilon_i, U_i), \quad i = 1, \ldots, 3$?
From Measurement system we can identify $\Sigma_\Theta$, $\Lambda$ subject to the usual normalizations.

(Look at the alternatives suggested by William’s notes; Carneiro et al.; Geweke, and Singleton; etc.)

Dedicated measurements case:

a 3 measurements per Factor (can relax with panels; see Cunha et al. 2008, 2010)

(This ignores the endogenous factors loading case)

$\therefore$ know $\Lambda$, $\Sigma_\Theta$, $\Sigma_V^M$
Appendix on Identification in Discrete Choice Models

(Econ 350, 2004)
Multinomial Probit Models
Also known as:

1. Thurstone Model V
2. Thurstone-Quandt Model
3. Developed by Domencich-McFadden (1978) (on reading list)

\begin{align*}
u_i &= v_i + \eta_i & i = 1, \ldots, J \\
v_i &= Z_i \beta \quad \text{(linear in parameters form)} \\
u_i &= Z_i \beta + \eta_i
\end{align*}
MNL
(i) $\beta$ fixed
(ii) $\eta_i$ iid

MNP
(i) $\beta$ random coefficient  $\beta \sim N(\bar{\beta}, \Sigma_\beta)$
(ii) $\beta$ independent of $\eta$  $\eta \sim (0, \Sigma_\eta)$,  
(allow gen. forms of correlation between errors)

\[ u_i = Z_i\bar{\beta} + Z_i (\beta - \bar{\beta}) + \eta_i \]

where $(\beta - \bar{\beta}) = \varepsilon$ and $Z_i (\beta - \bar{\beta}) + \eta_i$ is a composite heteroskedastic error term. ($\beta$ random=taste heterogeneity, $\eta_i$ can interpret as unobserved attributes of goods)
Main advantage of MNP over MNL is that it allows for general error covariance structure.

**Note:** To make computation easier, users sometimes set $\Sigma_\beta = 0$ (fixed coefficient version)
- allowing for $\beta$ random
- permits random taste variation
- allows for possibility that different persons value 2 characteristics differently
Introduction

We consider identification of parametric multinomial discrete choice models. This endeavor is – perhaps surprisingly – often analytically much harder than proving nonparametric identification of choice models.

For ease of exposition we begin by considering a multinomial probit model with three choices.
Let latent utility be described by

\[ U_1 = Z_1 \beta + \varepsilon_1, \]
\[ U_2 = Z_2 \beta + \varepsilon_2, \]
\[ U_3 = Z_3 \beta + \varepsilon_3, \]  \hspace{1cm} (4)

where \( Z \equiv (Z_1, Z_2, Z_3) \) are observable covariates assumed independent of \( \varepsilon \equiv (\varepsilon_1, \varepsilon_2, \varepsilon_3) \) and

\[ \varepsilon \sim \mathcal{N}(0, \Sigma). \]  \hspace{1cm} (5)
Consider first the probability of choosing alternative 1. Since only relative utilities are relevant rewrite (4) in difference form

\[ U_{1.2} = Z_{1.2} \beta + \varepsilon_{1.2}, \]
\[ U_{1.3} = Z_{1.3} \beta + \varepsilon_{1.3}, \]  

(6)

where the notation rule is \( X_{i,j} = X_i - X_j \).

Note that

\[ V(\varepsilon_{1.2}, \varepsilon_{1.3}) \equiv \Sigma_1 = \begin{bmatrix} \sigma_{11} + \sigma_{22} - 2\sigma_{12} & \sigma_{23} + \sigma_{11} - \sigma_{12} - \sigma_{13} \\ \sigma_{23} + \sigma_{11} - \sigma_{12} - \sigma_{13} & \sigma_{11} + \sigma_{33} - 2\sigma_{13} \end{bmatrix}. \]  

(7)
Let $D$ be the choice index equal to $j$ is alternative $j$ is chosen. Then

$$D = 1 \iff Z_{1.2\beta} + \varepsilon_{1.2} > 0, \quad Z_{1.3\beta} + \varepsilon_{1.3} > 0.$$ 

Since these inequalities are invariant to scale changes we can normalize by any positive constant. Dividing through by

$$\omega_{12} \equiv \sqrt{\sigma_{11} + \sigma_{22} - 2\sigma_{12}},$$

in the first inequality and

$$\omega_{13} \equiv \sqrt{\sigma_{11} + \sigma_{33} - 2\sigma_{13}},$$
in the second:

\[ D = 1 \iff Z_{1.2} \beta / \omega_{12} + \varepsilon_{1.2} / \omega_{12} > 0, \quad Z_{1.3} \beta / \omega_{13} + \varepsilon_{1.3} / \omega_{13} > 0, \]

and

\[ V(\varepsilon_{1.2} / \omega_{12}, \varepsilon_{1.3} / \omega_{13}) \equiv R_1 = \begin{bmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{bmatrix}, \quad (8) \]

where

\[ \rho_1 \equiv \frac{\sigma_{23} + \sigma_{11} - \sigma_{12} - \sigma_{13}}{(\sigma_{11} + \sigma_{22} - 2\sigma_{12})^{1/2}(\sigma_{11} + \sigma_{33} - 2\sigma_{13})^{1/2}}. \quad (9) \]
Then we can write the probability of choosing alternative 1 as

\[
\Pr(D = 1 \mid Z) = \int_{-\infty}^{Z_{1.2}/\omega_{12}} \int_{-\infty}^{Z_{1.3}/\omega_{13}} \phi(\tau \mid 0, R_1) \, d\tau_1 \, d\tau_2, \tag{10}
\]

where \( \phi(\tau \mid 0, R_1) \) is the bivariate normal density with mean zero and covariance matrix \( R_1 \).
By a parallel analysis

\[ Pr(D = 2|Z) = \int_{-\infty}^{Z_{2.1}\beta/\omega_{12}} \int_{-\infty}^{Z_{2.3}\beta/\omega_{23}} \phi(\tau|0, R_2) d\tau_1 d\tau_2, \quad (11) \]

where \( \omega_{23} \equiv \sqrt{\sigma_{22} + \sigma_{33} - 2\sigma_{23}} \) and

\[ R_2 = \begin{bmatrix} 1 & \rho_2 \\ \rho_2 & 1 \end{bmatrix}, \quad (12) \]

and

\[ \rho_2 \equiv \frac{\sigma_{22} - \sigma_{23} - \sigma_{12} + \sigma_{13}}{(\sigma_{11} + \sigma_{22} - 2\sigma_{12})^{1/2}(\sigma_{22} + \sigma_{33} - 2\sigma_{23})^{1/2}}. \quad (13) \]
Identification analysis

We now consider identification of the parameters of interest $\beta$ and $\Sigma$. We know the functions $\Pr(D = 1|Z)$ and $\Pr(D = 2|Z)$. Since $\Pr(D = 3|Z) = 1 - \Pr(D = 1|Z) - \Pr(D = 2|Z)$ this is all the information we have.

“Identification at infinity”

We first consider identification using an identification at infinity argument.
Consider $\Pr(D = 1|Z)$. If we can condition on very small values of $Z_3 \beta$ we can shut down alternative three as an option. In particular,

$$
\lim_{Z_3 \beta \to -\infty} \Pr(D = 1|Z) = \int_{-\infty}^{Z_{1.2} \beta / \omega_{12}} \phi(\tau|0, 1) d\tau,
$$

$$
= \Phi\left(Z_{1.2} \beta / \omega_{12}\right).
$$

(14)

Similarly, if we can condition on very small values of $Z_2 \beta$ we can shut down alternative two:

$$
\lim_{Z_2 \beta \to -\infty} \Pr(D = 1|Z) = \int_{-\infty}^{Z_{1.3} \beta / \omega_{13}} \phi(\tau|0, 1) d\tau,
$$

$$
= \Phi\left(Z_{1.3} \beta / \omega_{13}\right).
$$

(15)
By a similar argument we find

\[
\lim_{Z_1 \beta \to -\infty} \Pr(D = 2|Z) = \int_{-\infty}^{Z_{2.3\beta/\omega_{23}}} \phi(\tau|0, 1) d\tau,
\]

\[= \Phi\left(\frac{Z_{2.3\beta/\omega_{23}}}{\omega_{23}}\right), \tag{16}\]

\[
\lim_{Z_3 \beta \to -\infty} \Pr(D = 2|Z) = \int_{-\infty}^{Z_{2.1\beta/\omega_{12}}} \phi(\tau|0, 1) d\tau,
\]

\[= \Phi\left(\frac{Z_{2.1\beta/\omega_{12}}}{\omega_{12}}\right). \tag{17}\]
Using the classical results for identification of binary probit models we can therefore identify

\[ a = \beta/\omega_{12}, \]
\[ b = \beta/\omega_{13}, \]
\[ c = \beta/\omega_{23}. \]  \hspace{1cm} (18)

We now know the functions,

\[ P_1(\rho_1, Z) = \int_{-\infty}^{Z_{1.2}a} \int_{-\infty}^{Z_{1.3}b} \phi(\tau|0, R(\rho_1)) \, d\tau, \]  \hspace{1cm} (19)
\[ P_2(\rho_2, Z) = \int_{-\infty}^{Z_{2.1}a} \int_{-\infty}^{Z_{2.3}c} \phi(\tau|0, R(\rho_2)) \, d\tau. \]
Identification of $\rho_1$ and $\rho_2$ follows by appeal to the following lemma:

**Lemma**

*Fix some $\tilde{Z} \in \Omega_Z$. Then the mappings*

$$H_j(\rho_j) \equiv P_j(\rho_j, \tilde{Z}) : [-1, 1] \to [P_j, \overline{P_j}] \subseteq [0, 1], \quad j = 1, 2,$$

*are into.*
Proof. Consider the first mapping $H_1(\rho_1)$. We have

$$\frac{\partial H_1}{\partial \rho_1} = \int_{-\infty}^{\tilde{Z}_{1.2}} \int_{-\infty}^{\tilde{Z}_{1.3}} \frac{\partial \phi(\tau|0, R(\rho_1))}{\partial \rho_1} d\tau.$$
From Plackett (1954)\(^1\) we have the result:

\[
\frac{\partial H_1}{\partial \rho_1} = \int_{-\infty}^{\bar{Z}_{1.2}^a} \int_{-\infty}^{\bar{Z}_{1.3}^b} \frac{\partial \phi(\tau|0, R(\rho_1))}{\partial \rho_1} d\tau, \\
= \frac{\partial^2}{\partial (\bar{Z}_{1.2}^a) \partial (\bar{Z}_{1.3}^b)} \int_{-\infty}^{\bar{Z}_{1.2}^a} \int_{-\infty}^{\bar{Z}_{1.3}^b} \phi(\tau_1, \tau_2) d\tau_1 d\tau_2 \\
= \phi(\bar{Z}_{1.2}^a, \bar{Z}_{1.3}^b) > 0.
\]

Hence $\frac{\partial H_1}{\partial \rho_1} > 0$ and in addition we can show that $P_1 \leq H_1(\rho_1) \leq P_1$ for some $P_1 < P_1$. So the mapping $H_1(\rho_1)$ is into. A similar argument can be given for $H_2(\rho_2)$.

So the maps $H_1$ and $H_2$ are into, and we can solve $h_j = H_j(\rho_j)$ for a unique $\rho_1$ and $\rho_2$.

So we can identity $a, b, c$ and $\rho_1, \rho_2$. 
ON THE MEAN AND VARIANCE OF THE TETRACHORIC CORRELATION COEFFICIENT

MORTON B. BROWN
TEL-AVIV UNIVERSITY

JACQUELINE K. BENEDETTI
UNIVERSITY OF CALIFORNIA AT LOS ANGELES

Estimates of the mean and standard deviation of the tetrachoric correlation are compared with their expected values in several $2 \times 2$ tables. Significant bias in the mean is found when the minimum cell frequency is less than 5. Three formulas for the standard deviation are compared and guidelines given for their use.

Key words: fourfold correlation, $2 \times 2$ frequency table, bivariate normal.

This research was performed when the first author was on leave at the University of California at Los Angeles and was supported in part by NIH Special Research Resources Grant RR-3. The second author was also supported by NIH Fellowship 5 F22 GM00328-02.

Requests for reprints should be addressed to Dr. Jacqueline Benedetti at the Department of Biomathematics, AV-111, UCLA School of Medicine, Los Angeles, Ca. 90024.
Introduction

Data from a bivariate normal distribution are often displayed in a four-fold table where each variable is recorded as a dichotomy. Based upon the cell frequencies of the $2 \times 2$ table, Pearson [1901] proposed the tetrachoric correlation coefficient as an estimate of the correlation of the underlying bivariate normal distribution. The cell probabilities correspond to the probabilities of the four quadrants of the bivariate normal having the same marginals. For example, let the observed frequency table be

\[
\begin{array}{ccc}
  & a & b \\
 a + c & c + d & c + d \\
 a + b & b + d & N = a + b + c + d \\
\end{array}
\]

where $N$ is the total frequency.
Let $z_1$ and $z_2$ be standard normal deviates corresponding to the marginal probabilities $(a + c)/N$ and $(a + b)/N$; i.e.,

$$
\Phi(z_1) = \frac{(a + c)}{N}
$$

$$
\Phi(z_2) = \frac{(a + b)}{N}
$$

where $\Phi(z)$ is the cdf of the standard normal. Then the tetrachoric correlation $r$ is obtained by solving

$$
(1) \quad \int_{-\infty}^{z_2} \int_{-\infty}^{z_1} \phi(x_1, x_2, r) \, dx_1 \, dx_2 = \frac{a}{N}
$$

for $r$ where $\phi(x_1, x_2, r)$ is the bivariate normal density with means zero, variances one and correlation $r$. The probabilities of the four quadrants formed by dichotomizing the variables by the lines $x_1 = z_1$ and $x_2 = z_2$ are equal to $a/N$, $b/N$, $c/N$ and $d/N$. 
When the marginal probabilities are equal (i.e., \( z_1 = z_2 = 0 \)), the tetrachoric correlation \( r \) is obtained by

\[
(2) \quad r = -\cos \left( \frac{2\pi a}{N} \right).
\]

Otherwise the integral in (1) is replaced by the infinite tetrachoric series expansion in powers of \( r \) and a root of (1) must be found numerically. Manually this is a major undertaking and tables were prepared to ease some of the calculations [Everitt, 1910; Pearson, 1931]. Chesire et al. [1933] prepared computing diagrams to aid in finding \( r \). Recently computer programs have been written to solve the computational problem [e.g., see Froemel, 1971]. However, many programs err by truncating the infinite series too quickly.

Hamdan [1970] showed that the tetrachoric correlation is the maximum likelihood estimate of correlation in the \( 2 \times 2 \) table. Pearson [1901] also developed a formula for the standard error of the tetrachoric correlation which is tabulated by Hayes [1943] for several values of the correlation. Due to its computational complexity, Pearson [1913] proposed an approximation that could be tabulated in parts. Hamdan [1970] gives an alternate formula based on his maximum likelihood approach.
It is well-known that the product-moment correlation is a biased estimate of the underlying correlation. Similarly, we show numerically that the tetra-choric correlation is also biased, and indicate when bias is severe.

We also compute the population standard deviation and compare it with the formulas of Pearson and Hamdan. We then suggest guidelines for when these formulas can be used.
The Mean and Variance of the Tetrachoric Correlation

Let the underlying probabilities of the cells in a $2 \times 2$ contingency table be $(p_{11}, p_{12}, p_{21}, p_{22})$.

\[
\begin{array}{cc|c}
  p_{11} & p_{12} & p_{1+} \\
  p_{21} & p_{22} & p_{2+} \\
  \hline
  p_{+1} & p_{+2} & 1 \\
\end{array}
\]

As with Fisher's exact test, the marginal frequencies $p_{1+}$, $p_{2+}$, $p_{+1}$ and $p_{+2}$ are regarded as fixed a priori. The probability of observing the fourfold table $(a, b, c, d)$ is

\[
f(a, b, c, d) = k^{-1} \frac{p_{11}^a p_{12}^b p_{21}^c p_{22}^d}{a! b! c! d!}
\]

(3)

where

\[
k = \sum \frac{p_{11}^a p_{12}^b p_{21}^c p_{22}^d}{a! b! c! d!}
\]

subject to the constraints $a + b = Np_{1+}$ and $a + c = Np_{+1}$. Under the hypothesis that $\rho = 0$, (3) reduces to the hypergeometric density used in Fisher's exact test.
Let $\rho$ be the tetrachoric correlation of the fourfold table $(N_{p11}, N_{p12}, N_{p21}, N_{p22})$. Then the expected value of $r$ is
\begin{equation}
E(r) = \Sigma r f(a, b, c, d)
\end{equation}
and the expected mean square of $r$ is
\begin{equation}
\sigma^2(r) = \Sigma (r - \rho)^2 f(a, b, c, d)
\end{equation}
where the summations are subject to the constraints of (3). Note that $\sigma^2(r)$ is computed about $\rho$ and not about $E(r)$. 
Approximations of the Tetrachoric Correlation and Its Standard Error

The tetrachoric correlation is difficult to compute unless the marginal frequencies are equal (when the cosine function (2) can be used). Therefore many statistics have been proposed to approximate \( r \) [see Castellan, 1966]. A common approximation that uses the cosine function [Guilford, 1965] is

\[
(6) \quad r_{\text{cos} \cdot \text{pt}} = \cos! \left\{ \frac{\pi}{1 + \left( \frac{ad}{bc} \right)^{1/2}} \right\}
\]

which is exact when \( r = 0 \) or the marginal frequencies are equal (when it simplifies to (2)). We include it in our study to reemphasize the error which can occur when an approximation is used.
Pearson's [1901, 1913] formula for the standard deviation based on asymptotic theory is

\[ s_p = \frac{1}{N^{3/2}\phi(z_1, z_2, r)} \left\{ \frac{(a + d)(b + c)}{4} \right. \]
\[ + (a + c)(b + d)\Phi_2^2 + (a + b)(c + d)\Phi_1^2 \]
\[ + 2(ad - bc)\Phi_1\Phi_2 - (ab - cd)\Phi_2 - (ac - bd)\Phi_1 \right\}^{1/2} \]

where

\[ \Phi_1 = \Phi \left( \frac{z_1 - rz_2}{(1 - r^2)^{1/2}} \right) - 0.5 \]
\[ \Phi_2 = \Phi \left( \frac{z_2 - rz_1}{(1 - r^2)^{1/2}} \right) - 0.5 \]

and

\[ \phi(z_1, z_2, r) = \frac{1}{2\pi(1 - r^2)^{1/2}} \exp \left\{ - \frac{z_1^2 - 2rz_1z_2 + z_2^2}{2(1 - r^2)} \right\}. \]
(Note that Pearson defines probable error, which is $0.67449s_p$.) Due to the complex hand calculations for this formula he gives an approximation to it [Pearson, 1913]:

$$s_A = \frac{\left(\frac{(a + b)(a + c)(b + d)(c + d)}{N^5}\right)(1 - r^2)\left[1 - \left(\frac{\sin^{-1} r}{\pi/2}\right)^2\right]}{\phi(z_1, z_2, 0)}$$

(8)

[see also Guilford and Lyons, 1942].

Based upon maximum likelihood, Hamdan [1970] derives an asymptotic standard error as

$$s_H = \frac{1}{N\phi(z_1, z_2, r)}\left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right)^{-1/2}.$$  

(9)

Note that when $r = 0$, $s_p$, $s_A$ and $s_H$ all reduce to

$$s_0 = \frac{\left\{\frac{(a + b)(a + c)(b + d)(c + d)}{N^5}\right\}^{1/2}}{\phi(z_1, z_2, 0)},$$

(10)

which is appropriate to test the null hypothesis that the correlation is zero.
Results

The parameters $\rho$, $E(r)$ and $\sigma(r)$ and the statistics $r_{cos.pl}$, $s_p$, $s_H$, $s_A$, $s_0$ are computed for various tables ($Np_{11}$, $Np_{12}$, $Np_{21}$, $Np_{22}$). The evaluation of $r$ for each “observed” table ($a$, $b$, $c$, $d$) in (4) and (5) is by a subroutine that uses a varying number of terms in the tetrachoric series such that the bivariate normal probability of a quadrant is evaluated with an error less than $2 \times 10^{-6}$.

When $\rho = 0$, the results are given in Table 1 and when $\rho \neq 0$ they are in Table 2.
<table>
<thead>
<tr>
<th>P_{11}</th>
<th>P_{12}</th>
<th>P_{21}</th>
<th>P_{22}</th>
<th>N</th>
<th>min. cell frequency</th>
<th>E(r)</th>
<th>σ(r)</th>
<th>s_0 = s_P = s_H = s_A</th>
</tr>
</thead>
<tbody>
<tr>
<td>.25</td>
<td>.25</td>
<td>.25</td>
<td>.25</td>
<td>40</td>
<td>10</td>
<td>0</td>
<td>.2439</td>
<td>.2484</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>100</td>
<td>25</td>
<td>0</td>
<td>.1559</td>
<td>.1571</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>400</td>
<td>100</td>
<td>0</td>
<td>.0784</td>
<td>.0785</td>
</tr>
<tr>
<td>.05</td>
<td>.05</td>
<td>.45</td>
<td>.45</td>
<td>40</td>
<td>2</td>
<td>0</td>
<td>.4144</td>
<td>.3387</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>100</td>
<td>5</td>
<td>0</td>
<td>.2206</td>
<td>.2142</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>400</td>
<td>20</td>
<td>0</td>
<td>.1077</td>
<td>.1071</td>
</tr>
<tr>
<td>.01</td>
<td>.01</td>
<td>.49</td>
<td>.49</td>
<td>100</td>
<td>1</td>
<td>0</td>
<td>.7035</td>
<td>.3624</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>400</td>
<td>4</td>
<td>0</td>
<td>.2045</td>
<td>.1812</td>
</tr>
<tr>
<td>.01</td>
<td>.09</td>
<td>.09</td>
<td>.81</td>
<td>100</td>
<td>1</td>
<td>-.2481</td>
<td>5.9990</td>
<td>.2922</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>400</td>
<td>4</td>
<td>-.0257</td>
<td>1.825</td>
<td>.1461</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1000</td>
<td>10</td>
<td>-.0074</td>
<td>.0951</td>
<td>.0924</td>
</tr>
<tr>
<td>.002</td>
<td>.018</td>
<td>.098</td>
<td>.882</td>
<td>500</td>
<td>1</td>
<td>-.2851</td>
<td>.6002</td>
<td>.2210</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2500</td>
<td>5</td>
<td>-.0175</td>
<td>.1250</td>
<td>.0989</td>
</tr>
<tr>
<td>.0004</td>
<td>.0196</td>
<td>.0196</td>
<td>.9604</td>
<td>2500</td>
<td>1</td>
<td>-.3178</td>
<td>.6067</td>
<td>.1672</td>
</tr>
</tbody>
</table>
TABLE 2
Tables for which the Tetrachoric Coefficient is Not Zero

<table>
<thead>
<tr>
<th>$p_{11}$</th>
<th>$p_{12}$</th>
<th>$p_{21}$</th>
<th>$p_{22}$</th>
<th>$N$</th>
<th>min. cell frequency</th>
<th>$r$</th>
<th>$E(r)$</th>
<th>$r_{cos.pi}$</th>
<th>$\sigma(r)$</th>
<th>$s_p$</th>
<th>$s_H$</th>
<th>$s_A$</th>
<th>$s_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>.3333</td>
<td>.1667</td>
<td>.1667</td>
<td>.3333</td>
<td>30</td>
<td>5</td>
<td>.5000</td>
<td>.4968</td>
<td>.5000</td>
<td>.2308</td>
<td>.2342</td>
<td>.2342</td>
<td>.2342</td>
<td>.2868</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>90</td>
<td>15</td>
<td>.4989</td>
<td>.4997</td>
<td></td>
<td>.1346</td>
<td>.1352</td>
<td>.1352</td>
<td>.1352</td>
<td>.1656</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>300</td>
<td>50</td>
<td>.0739</td>
<td>.0740</td>
<td>.0740</td>
<td>.0740</td>
<td>.0740</td>
<td>.0740</td>
<td>.0740</td>
<td>.0907</td>
</tr>
<tr>
<td>.45</td>
<td>.05</td>
<td>.05</td>
<td>.45</td>
<td>40</td>
<td>2</td>
<td>.9511</td>
<td>.9501</td>
<td>.9511</td>
<td>.0482</td>
<td>.0460</td>
<td>.0460</td>
<td>.0460</td>
<td>.2484</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>100</td>
<td>5</td>
<td>.9507</td>
<td>.9507</td>
<td></td>
<td>.0297</td>
<td>.0291</td>
<td>.0291</td>
<td>.0291</td>
<td>.1571</td>
</tr>
<tr>
<td>.08</td>
<td>.02</td>
<td>.42</td>
<td>.48</td>
<td>100</td>
<td>2</td>
<td>.4395</td>
<td>.4742</td>
<td>.5394</td>
<td>.2398</td>
<td>.1919</td>
<td>.1908</td>
<td>.1842</td>
<td>.2142</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>400</td>
<td>8</td>
<td>.4432</td>
<td>.4432</td>
<td></td>
<td>.0970</td>
<td>.0960</td>
<td>.0954</td>
<td>.0921</td>
<td>.1071</td>
</tr>
<tr>
<td>.0967</td>
<td>.0033</td>
<td>.4033</td>
<td>.4967</td>
<td>300</td>
<td>1</td>
<td>.7483</td>
<td>.8132</td>
<td>.9003</td>
<td>.1611</td>
<td>.0890</td>
<td>.0877</td>
<td>.0692</td>
<td>.1237</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>600</td>
<td>2</td>
<td>.7747</td>
<td>.7747</td>
<td></td>
<td>.1026</td>
<td>.0630</td>
<td>.0620</td>
<td>.0489</td>
<td>.0875</td>
</tr>
<tr>
<td>.002</td>
<td>.018</td>
<td>.498</td>
<td>.482</td>
<td>500</td>
<td>1</td>
<td>-.4733</td>
<td>-.6097</td>
<td>-.7138</td>
<td>.3261</td>
<td>.1594</td>
<td>.1585</td>
<td>.1355</td>
<td>.1621</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1000</td>
<td>2</td>
<td>-.5267</td>
<td>-.5267</td>
<td></td>
<td>.2030</td>
<td>.1127</td>
<td>.1121</td>
<td>.0958</td>
<td>.1146</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1500</td>
<td>3</td>
<td>-.4981</td>
<td>-.4981</td>
<td></td>
<td>.1371</td>
<td>.0920</td>
<td>.0936</td>
<td>.0915</td>
<td>.0783</td>
</tr>
<tr>
<td>.075</td>
<td>.025</td>
<td>.025</td>
<td>.875</td>
<td>40</td>
<td>1</td>
<td>.9358</td>
<td>.9292</td>
<td>.9612</td>
<td>.0968</td>
<td>.0851</td>
<td>.0842</td>
<td>.1038</td>
<td>.4620</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>80</td>
<td>2</td>
<td>.9328</td>
<td>.9328</td>
<td></td>
<td>.0644</td>
<td>.0602</td>
<td>.0595</td>
<td>.0734</td>
<td>.3267</td>
</tr>
<tr>
<td>.03</td>
<td>.07</td>
<td>.07</td>
<td>.83</td>
<td>100</td>
<td>3</td>
<td>.4599</td>
<td>.4220</td>
<td>.5691</td>
<td>.2779</td>
<td>.2175</td>
<td>.2163</td>
<td>.2472</td>
<td>.2922</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>400</td>
<td>12</td>
<td>.4535</td>
<td>.4535</td>
<td></td>
<td>.1104</td>
<td>.1087</td>
<td>.1082</td>
<td>.1236</td>
<td>.1461</td>
</tr>
<tr>
<td>.015</td>
<td>.085</td>
<td>.485</td>
<td>.415</td>
<td>400</td>
<td>6</td>
<td>-.5197</td>
<td>-.5249</td>
<td>-.6378</td>
<td>.0939</td>
<td>.0915</td>
<td>.0908</td>
<td>.0858</td>
<td>.1071</td>
</tr>
</tbody>
</table>
Bias of the Tetrachoric Correlation

When $\rho = 0$ and $p_{1+}$ and/or $p_{+1}$ is equal to $\frac{1}{2}$, $E(r) = 0$. Otherwise $E(r)$ is a biased estimate of $\rho$. The bias is most severe when one of the cells has an expected frequency near zero. When no frequency is less than 5, the bias is negligible.

This is easily understood by considering the change in $r$ as any observed cell frequency approaches zero. For example, the table $(1, 9, 9, 81)$ yields $r = 0$, whereas for $(0, 10, 10, 80)$, $r = -1.0$. When the theoretical probabilities are $(0.01, 0.09, 0.09, 0.81)$ and $N = 100$, the probability that $a = 0$ (and $r = 1.0$) is large. Hence the expected value is strongly influenced by a zero cell. When $N = 400$ in the above, the probability that $a = 0$ is greatly reduced, as is the bias.

Therefore caution is necessary in interpreting $r$ especially when any observed frequency is less than 5.
When One Cell is Zero

When one and only one cell is zero, it is obvious that $r = \pm 1.0$ for that table. However, the computation of $r$ is based upon the premise that the observed frequencies represent the underlying frequencies exactly. It is clear that the observed frequencies are integer quantities and therefore cannot be more accurate than $\frac{1}{2}$ unit. Hence the cell probability can be no more accurate than $1/(2N)$.

We therefore replaced the zero observed frequency by $\frac{1}{2}$ and modified the other cells to maintain the same marginal totals. When the expected value $E(r)$ was recomputed using this modification, the bias was greatly reduced. For example when $N = 100$, the table with probabilities $(0.01, 0.09, 0.09, 0.81)$ had a modified expected value 0.0187 instead of $-0.2481$. Similar large reductions were obtained whenever there was an expected frequency of 1 in the table.

The tetrachoric correlation obtained by modifying the zero cell to $\frac{1}{2}$ is the minimum $r$ which is consistent with the rounding errors inherent in using the observed frequencies. Therefore we recommend this adjustment (similar to Yates correction) whenever a single cell is zero.

When two cells are zero (either diagonal or off-diagonal), one of the two variables is redundant and we would assign a correlation of $\pm 1.0$. 
The Error in $r_{\cos.pl}$

When $ad = bc$ both $r$ and $r_{\cos.pl}$ are zero. When all the marginals are equal, $r$ (which is obtained by (2)) is equivalent to $r_{\cos.pl}$. Otherwise $r_{\cos.pl}$ may greatly overestimate the correlation. This can be seen in Table 2 by comparing $\rho$ with its estimate $r_{\cos.pl}$. Therefore, we recommend that $r_{\cos.pl}$ not be used.
The Behavior of the Standard Errors

From Table 1 we see that $s_0$ rapidly converges to $\sigma(r)$ when $r = 0$. When the minimum cell frequency is at least 5, the convergence is adequate. Note that $s_0 = s_p = s_H = s_A$ when $r = 0$.

When $r \neq 0$ (Table 2), the standard errors $s_p$ and $s_H$ converge slowly to $\sigma(r)$. Pearson's approximate formula $s_A$, whose development was based on computational ease, is the most biased. Since Hamdan's standard error $s_H$ is similar to $s_p$ and easiest to calculate of the three formulas, we recommend its use. The convergence of $s_p$ and $s_H$ to $\sigma(r)$ when $\rho \neq 0$ is slower than when $\rho = 0$. Therefore we recommend their use when the minimum cell frequency is at least 10, unless the marginal totals are equal when the minimum cell frequency need only be 5.
The Computation of the Tetrachoric Correlation

Froemel [1971] provides a fine comparison of three routines that are available to compute the tetrachoric coefficient. Others are continually being written. Most routines are based upon fitting a tetrachoric series expansion in $r$ and then finding a root of the expansion. For example, the routine in the IBM Scientific Subroutine Package [1970] uses the first six terms. Martinson and Hamdan [1975] use the first eight terms in finding the polychoric estimate of correlation which is the tetrachoric correlation when the table is $2 \times 2$.

Using a criterion that the tetrachoric series has converged when two successive terms are less than $10^{-6}$, we found the following approximate relationship between the number of terms and $r$.

| $|r|$ | 0.25 | 0.5 | 0.75 | 0.8 | 0.85 | 0.9 | 0.95 |
|-----|------|-----|------|-----|------|-----|------|
| terms | 10   | 18  | 36   | 48  | 60   | 86  | 150  |

Even if the criterion is less rigid, the number of terms must increase with $|r|$. Therefore any routine using a short fixed length series will not be accurate when $|r|$ is large.
REFERENCES


Pearson, K. On the probable error of a coefficient of correlation as found from a fourfold table. *Biometrika*, 1913, 9, 22–27.


*Manuscript received 6/25/76*

*Final version received 12/10/76*
Specific normalizations

(I have additional normalization in the main text that use the outcome and measurement structure.)

Normalization 1

This normalization has $\beta$ free and

$$\Sigma = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \sigma_{22} & 0 \\ 0 & 0 & \sigma_{33} \end{bmatrix}.$$
We know the left hand side of the following equation system:

\[
\begin{align*}
a &= \frac{\beta}{\sqrt{1 + \sigma_{22}}} \\
b &= \frac{\beta}{\sqrt{1 + \sigma_{33}}} \\
c &= \frac{\beta}{\sqrt{\sigma_{22} + \sigma_{33}}} \\
\rho_1 &= \frac{1}{\sqrt{1 + \sigma_{22}} \sqrt{1 + \sigma_{33}}} \\
\rho_2 &= \frac{\sigma_{22}}{\sqrt{1 + \sigma_{22}} \sqrt{\sigma_{22} + \sigma_{33}}}
\end{align*}
\]

(20)
Then since
\[ \frac{b}{a} = \left[ \frac{1 + \sigma_{22}}{1 + \sigma_{33}} \right]^{1/2}, \]
we have
\[ \rho_1 = \frac{b}{a} \frac{1}{1 + \sigma_{22}}. \]
So we can solve for \( \sigma_{22} \),
\[ \sigma_{22} = \frac{b}{a \rho_1} - 1, \quad \rho_1 \neq 0. \] (21)
Knowing \( \sigma_{22} \) we can solve immediately for \( \beta \) and \( \sigma_{33} \).
With only three free parameters \((\beta, \sigma_{22}, \sigma_{33})\) and five equations this normalization seems strongly overidentified. Yet removing the constraint \(\sigma_{11} = 1\) one finds that \((\beta, \sigma_{11}, \sigma_{22}, \sigma_{33})\) are not identified.
Normalization II

This normalization has $\beta$ free and

$$\Sigma = \begin{bmatrix} 1 & \sigma_{12} & 0 \\ \sigma_{12} & \sigma_{22} & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$
We know the left hand side of the following equation system:

\[
\begin{align*}
    a &= \frac{\beta}{\sqrt{1 + \sigma_{22} - 2\sigma_{12}}} \\
    b &= \beta \\
    c &= \frac{\beta}{\sqrt{\sigma_{22}}} \\
    \rho_1 &= \frac{1 - \sigma_{12}}{\sqrt{1 + \sigma_{22} - 2\sigma_{12}}} \\
    \rho_2 &= \frac{\sigma_{22} - \sigma_{12}}{\sqrt{1 + \sigma_{22} - 2\sigma_{12}} \sqrt{\sigma_{22}}}
\end{align*}
\]  (22)
Clearly $\beta$ is immediately identified from $b$ and we then get $\sigma_{22}$ from $c$. Finally, we can solve $\sigma_{12}$ from the relation,

$$\left(\frac{a}{c}\right)^2 = \frac{\sigma_{22}}{1 + \sigma_{22} - 2\sigma_{12}}.$$
Normalization III

This normalization has $\beta$ free and

$$\Sigma = \begin{bmatrix} 1 & \alpha_1 & \alpha_1 \alpha_2 & \alpha_1 \alpha_3 \\ \alpha_1 & 1 & \alpha_2 & \alpha_2 \alpha_3 \\ \alpha_1 \alpha_2 & \alpha_2 & 1 & \alpha_3 \\ \alpha_1 \alpha_3 & \alpha_2 \alpha_3 & 1 & \end{bmatrix}.$$ 

The restrictions on $\Sigma$ are the restrictions implied by a one-factor model for $\varepsilon$:

$$\varepsilon_j = \alpha_j f + \mu_j,$$

with $f \sim N(0, 1)$ and $V[\varepsilon] = 1$. 
We know the left hand side of the following equation system:

\[
\begin{align*}
da &= \frac{\beta/\sqrt(2)}{(1 - \alpha_1 \alpha_2)^{1/2}} \\
b &= \frac{\beta/\sqrt(2)}{(1 - \alpha_1 \alpha_3)^{1/2}} \\
c &= \frac{\beta/\sqrt(2)}{(1 - \alpha_2 \alpha_3)^{1/2}} \\
\rho_1 &= \frac{(1/2)(1 + \alpha_2 \alpha_3 - \alpha_1 \alpha_2 - \alpha_1 \alpha_3)}{(1 - \alpha_1 \alpha_2)^{1/2}(1 - \alpha_1 \alpha_3)^{1/2}} \\
\rho_2 &= \frac{(1/2)(1 + \alpha_1 \alpha_3 - \alpha_1 \alpha_2 - \alpha_2 \alpha_3)}{(1 - \alpha_1 \alpha_2)^{1/2}(1 - \alpha_2 \alpha_3)^{1/2}}
\end{align*}
\] (23)
Then

\[
\left(\frac{a}{c}\right)^2 = \frac{1 - \alpha_2 \alpha_3}{1 - \alpha_1 \alpha_2},
\]
\[
\left(\frac{b}{c}\right)^2 = \frac{1 - \alpha_2 \alpha_3}{1 - \alpha_1 \alpha_3}.
\]

Solving for \(\alpha_1 \alpha_2\) and \(\alpha_1 \alpha_3\) we get

\[
\alpha_1 \alpha_2 = \frac{(a/d)^2 + \alpha_2 \alpha_3 - 1}{(a/d)^2} \equiv k_1 + k_2 \alpha_2 \alpha_3, \tag{24}
\]
\[
\alpha_1 \alpha_3 = \frac{(b/d)^2 + \alpha_2 \alpha_3 - 1}{(b/d)^2} \equiv m_1 + m_2 \alpha_2 \alpha_3, \tag{25}
\]

where \(k_1 = 1 - (d/a)^2\), \(k_2 = (d/a)^2\) and \(m_1 = 1 - (d/b)^2\), \(m_2 = (d/b)^2\).
Substituting this in the expression for $\rho_2$ we find

$$2\rho_2 = \frac{1 + m_1 - k_1 + (m_2 - k_2 - 1)\alpha_2\alpha_3}{(1 - k_1 - k_2\alpha_2\alpha_3)^{1/2}(1 - \alpha_2\alpha_3)^{1/2}}.$$ 

Everything but $\alpha_2\alpha_3$ is known in this equation. This can be rewritten as the quadratic equation

$$(\alpha_2\alpha_3)^2 + q_1\alpha_2\alpha_3 + q_2 = 0,$$

with $q_1$ and $q_2$ appropriately defined as functions of $k_1, k_2, m_1, m_2$.

Assuming away complex solutions we can thus determine $\alpha_2\alpha_3$ up to sign. From (24) and (25) we can then determine $\alpha_1\alpha_2$ and $\alpha_1\alpha_3$. Identification of $\beta$ follows immediately.
Problem of Identification and Normalization in the MNP model

Reference:

David Bunch (1979), “Estimability in the multinominal Probit Model” in *Transportation Research*

Domencich and McFadden
Let \( Z \bar{\beta} = \begin{pmatrix} Z_1 \cdot \bar{\beta} \\ \vdots \\ Z_J \cdot \bar{\beta} \end{pmatrix} \) and \( \tilde{\eta} = \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_J \end{pmatrix} \) for \( J \) alternatives and \( K \) characteristics. \( \beta \) is random with \( \beta \sim N(\bar{\beta}, \Sigma_{\beta}) \).

\[
\Pr \text{ (alternative } j \text{ selected)} = \Pr (u_j > u_i) \quad \forall i \neq j
\]

\[
= \int_{u_j = -\infty}^{\infty} \int_{u_i = -\infty}^{u_j} \int_{u_J = -\infty}^{u_j} \Phi(u \mid V_\mu, \Sigma_\mu) \, du_J \, du_I \, du_j
\]

where \( \Phi(u \mid V_\mu, \Sigma_\mu) \) is pdf.

\( (\Phi \) is \( J \)-dimensional MVN density with mean \( V_\mu, \Sigma_\mu \))

**Note:** Unlike the MVL, no closed form expression for the integral. The integrals often evaluated using simulation methods (we will work an example).
How many parameters are there?

\[ \bar{\beta} : \quad K \text{ parameters} \]
\[ \Sigma_{\beta} : \quad K \times K \text{ symmetric matrix} \quad \frac{K^2 - K}{2} + K = \frac{K(K + 1)}{2} \]
\[ \Sigma_{\eta} : \quad \frac{J(J + 1)}{2} \]

**Note:** When a person chooses \( j \), all we know is relative utility, not absolute utility. This suggests that not all parameters in the model will be identified. Will require normalizations.
Digression on Identification

What does it mean to say a parameter is not identified in a model?

⇒ model with one parameterization is observationally equivalent to another model with a different parameterization
Example: Binary Probit Model (fixed $\beta$)

$$
\Pr(D = 1 \mid Z) = \Pr(v_1 + \varepsilon_1 > v_2 + \varepsilon_2) \\
= \Pr(x\beta + \varepsilon_1 > x_2\beta + \varepsilon_2) \\
= \Pr((x_1 - x_2)\beta > \varepsilon_2 - \varepsilon_1) \\
= \Pr\left(\frac{(x_1 - x_2)\beta}{\sigma} > \frac{\varepsilon_2 - \varepsilon_1}{\sigma}\right) \\
= \Phi\left(\frac{\tilde{x}\beta}{\sigma}\right), \quad \tilde{x} = x_1 - x_2
$$

$\Phi\left(\frac{\tilde{x}\beta}{\sigma}\right)$ is observationally equivalent to $\Phi\left(\frac{\tilde{x}\beta^*}{\sigma^*}\right)$ for $\frac{\beta}{\sigma} = \frac{\beta^*}{\sigma^*}$. 
\( \beta \) not separably identified relative to \( \sigma \) but ratio is identified:

\[
\begin{align*}
\Phi \left( \frac{\tilde{x}\beta}{\sigma} \right) & = \Phi \left( \frac{\tilde{x}\beta^*}{\sigma^*} \right) \\
\Phi^{-1} \cdot \Phi \left( \frac{\tilde{x}\beta}{\sigma} \right) & = \Phi^{-1} \Phi \left( \frac{\tilde{x}\beta^*}{\sigma^*} \right) \\
\Rightarrow \quad \frac{\beta}{\sigma} & = \frac{\beta^*}{\sigma^*}
\end{align*}
\]

Set \( \{ b : b = \beta \cdot \delta, \ \delta \text{ any positive scalar} \} \) is identified (say “\( \beta \) is identified up to scale and sign is identified”).
Identification in the MVP model

$$\Pr (j \text{ selected} \mid V_\mu, \Sigma_\mu) = \Pr (u_i - u_j < 0 \quad \forall i \neq j)$$

Define $$\Delta_j = \begin{pmatrix} 1 & 0 & .. & -1 & .. & 0 \\ 0 & 1 & .. & -1 & .. & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & .. & .. & -1 & 0 & 1 \end{pmatrix} (J-1) \times J$$ (contrast matrix)

$$\Delta_j \tilde{u} = \begin{pmatrix} u' - u^j \\ \vdots \\ u^j - u^j \end{pmatrix}$$
\[
\Pr (j \text{ selected} \mid V_\mu, \Sigma_\mu) = \Pr (\Delta_j \tilde{u} < 0 \mid V_\mu, \Sigma_\mu) = \Phi (0 \mid V_Z, \Sigma_Z)
\]

where

(i) \(V_Z\) is the mean of \(\Delta_j \tilde{u} = \Delta_j \tilde{Z} \bar{\beta}\)

(ii) \(\Sigma_Z\) is the variance of \(\Delta_j \tilde{Z} \Sigma_\beta \tilde{Z}' \Delta_j' + \Delta_j \Sigma_\eta \Delta_j'

(iii) \(V_Z\) is \((J - 1) \times 1\)

(iv) \(\Sigma_Z\) is \((J - 1) \times (J - 1)\)

⇒ We reduce dimensions of the integral by one.
This says that all of the information exists in the contrasts. Can’t identify all the components because we only observe the contrasts. Now define $\tilde{\Delta}_j$ as $\Delta_j$ with $J$th column removed and choose $J$ as the reference alternative with corresponding $\Delta_J$. Then can verify that

$$\Delta_j = \tilde{\Delta}_j \cdot \Delta_J$$
For example, with three goods:

\[
\begin{pmatrix}
1 & -1 \\
0 & -1
\end{pmatrix}
×
\begin{pmatrix}
1 & 0 & -1 \\
0 & 1 & 1
\end{pmatrix}
= 
\begin{pmatrix}
1 & -1 & 0 \\
0 & -1 & 1
\end{pmatrix}
\]

\( \tilde{\Delta}_j, (j = 2), \quad \Delta_j, (J = 3), \quad \Delta_j, (j = 2, 3rd \text{ column included}) \)

3rd column removed, reference alt.)
Therefore, we can write

\[ V_Z = \Delta_j \tilde{Z} \tilde{\beta} \]
\[ \Sigma_Z = \Delta_j \tilde{Z} \Sigma_\beta \tilde{Z}' \Delta_j' + \tilde{\Delta}_j \Delta_J \Sigma_\eta \Delta'_j \tilde{\Delta}'_j \]

where \( C_J = \Delta_J \Sigma_\eta \Delta'_J \) and \((J - 1) \times (J - 1)\) has
\[
\frac{(J - 1)^2}{2} - (J - 1) + (J + 1) \text{ parameters} = \frac{J(J - 1)}{2} \text{ total. Since original model can always be expressed in terms of a model with } \( (\beta, \Sigma_\beta, C_J) \), it follows that some of the parameters in the original model are not identified.
How many parameters not identified?

Original model:

\[ K + \frac{K(K + 1)}{2} + \frac{J(J + 1)}{2} \]

Now:

\[ K + \frac{K(K + 1)}{2} + \frac{J(J - 1)}{2}, \quad \frac{J^2 + J - (J^2 - J)}{2} = J \text{ not identified} \]

Turns out that one additional parameter not identified.

Total: \( J + 1 \)

**Note:** Evaluation of \( \Phi(0 \mid kv_Z, k^2\Sigma_Z) \) \( k > 0 \) gives same result as evaluating \( \Phi(0 \mid v_Z, \Sigma_Z) \) can eliminate one more parameter by suitable choice of \( k \).
Illustration

\[ J = 3 \quad \Sigma_\eta = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{pmatrix} \]

\[ C_2 = \Delta_2 \Sigma_\eta \Delta'_2 = \begin{pmatrix} 1 & -1 & 0 \\ 0 & -1 & 1 \end{pmatrix} \cdot \Sigma_\eta \begin{pmatrix} 1 & -1 & 0 \\ 0 & -1 & 1 \end{pmatrix}' \]

\[ = \begin{pmatrix} \sigma_{11} & -2\sigma_{21} & +\sigma_{22}, \quad \sigma_{21} & -\sigma_{31} & -\sigma_{32} & +\sigma_{22} \\ \sigma_{21} & -\sigma_{31} & -\sigma_{32} & +\sigma_{22}, \quad \sigma_{33} & -2\sigma_{31} & +\sigma_{22} \end{pmatrix} \]

\[ C_2 = \tilde{\Delta}_2 \Delta_3 \Sigma_\eta \Delta'_3 \Delta'_2 = \begin{pmatrix} 1 & -1 \\ 0 & -1 \end{pmatrix} \cdot \begin{pmatrix} \sigma_{11} & -2\sigma_{21} & +\sigma_{33}, \quad \sigma_{21} & -\sigma_{31} & -\sigma_{32} & +\sigma_{33} \\ \sigma_{21} & -\sigma_{31} & -\sigma_{32} & +\sigma_{33} \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ -1 & -1 \end{pmatrix} \]
Normalization Approach of Albreit, Lerman, and Manski (1978)

**Note:** Need $J + 1$ restrictions on VCV matrix.

- fix $J$ parameters by setting last row and last column of $\Sigma_\eta$ to 0
- fix scale by constraining diagonal elements of $\Sigma_\eta$ so that $\frac{\sum_\varepsilon}{J}$ equals variance of a standard Weibull. (To compare estimates with MNL and independent probit)
How do we solve the forecasting problem?

Suppose that we have 2 goods and add a 3rd

\[
\Pr(1 \text{ chosen}) = \Pr(u^1 - u^2 \geq 0) = \Pr(1 \left( (Z^1 - Z^2) \bar{\beta} \geq \omega^2 - \omega^1 \right)
\]

where

\[
\omega^1 = Z^1 (\beta - \bar{\beta}) + \eta^1, \quad \omega^2 = Z^2 (\beta - \bar{\beta}) + \eta^2
\]

\[
= \int_{-\infty}^{\infty} \left( \frac{(Z^1 - Z^2)\bar{\beta}}{\sigma_{11} + \sigma_{22} - 2\sigma_{12} + (Z^2 - Z^1)' \Sigma \eta (Z^2 - Z_1)'} \right)^{1/2} \frac{1}{\sqrt{2\pi}} e^{-t/2} dt
\]

Now add a 3rd good

\[
u^3 = Z^3 \bar{\beta} + Z^3 (\beta - \bar{\beta}) + \eta^3.
\]
Problem: We don’t know correlation of $\eta^3$ with other errors.

Suppose that $\eta^3 = 0$ (i.e. only preference heterogeneity). Then

$$Pr(1 \text{ chosen}) = \int_{-\infty}^{a} \int_{-\infty}^{b} B.V.N. \, dt_1 dt_2$$

when

$$a = \frac{(Z^1 - Z^2) \bar{\beta}}{\left[\sigma_{11} + \sigma_{22} - 2\sigma_{12} + (Z^2 - Z^1) \Sigma_\beta (Z^2 - Z^1)'\right]^{1/2}}$$

and

$$b = \frac{(Z^1 - Z^3) \bar{\beta}}{\left[\sigma_{11} + (Z^3 - Z^1) \Sigma_\beta (Z^3 - Z^1)'\right]^{1/2}}$$
We could also solve the forecasting problem if we make an assumption like \( \eta^2 = \eta^3 \).

We solve red-bus//blue-bus problem if \( \eta^2 = \eta^1 = 0 \) and \( z^3 = z^2 \).

\[
\Pr (1 \text{ chosen}) = \Pr (u^1 - u^2 \geq 0, u^1 - u^3 \geq 0)
\]

but \( u^1 - u^2 \geq 0 \land u^1 - u^3 \geq 0 \) are the same event.

\( \therefore \) adding a third choice does not change the choice of 1.
Estimation Methods for MNP Models

Models tend to be difficult to estimate because of high dimensional integrals. Integrals need to be evaluated at each stage of estimating the likelihood.

Simulation provides a means of estimating $P_{ij} = \Pr (i \text{ chooses } j)$
1. We suppose that the vector $\mathbf{x}$ of $n$ random variables with zero means and unit variances has a non-singular normal multivariate distribution with probability density function

$$
\phi_n(x_1, x_2, \ldots, x_n; c_{11}, c_{12}, \ldots, c_{nn}) = (2\pi)^{-\frac{n}{2}} |C|^{\frac{1}{2}} \exp\left(-\frac{1}{2} \mathbf{x}' \mathbf{C} \mathbf{x}\right).
$$

(1)

Here $C$ is the inverse of the variance matrix of $\mathbf{x}$, which is denoted by $\mathbf{R}$, and has elements $\{\rho_{ij}\}$. The integral

$$
\Phi_n(a_1, a_2, \ldots, a_n; c_{11}, c_{12}, \ldots, c_{nn}) = \int_{a_1}^{\infty} \int_{a_2}^{\infty} \cdots \int_{a_n}^{\infty} \phi_n \, dx_1 \, dx_2 \cdots dx_n
$$

(2)

has been extensively tabulated for $n$ equal to 1 and 2, but a simple calculation will show that comparable tables are scarcely feasible for higher values of $n$. However, the joint distribution of several continuous dependent variables is often assumed to conform to multivariate normality, and experiments may arise in which we wish to compare observed and expected frequencies, as, for example, in testing models for the joint action of several poisons (Plackett & Hewlett, 1952).
When \( a_1, a_2, \ldots, a_n \) are all zero, we denote the value of \( \Phi_n \) by \( \Phi_n^0 \); it can be interpreted as the area of a general simplex constructed on the surface of a sphere in \( n \) dimensions. Schläfli (1858) found a differential recurrence relation for \( \Phi_n^0 \), and showed that, when \( n \) is odd, \( \Phi_n^0 \) can be expressed as a combination of values of \( \Phi_{n-1}^0, \Phi_{n-3}^0, \ldots, \Phi_2^0 \). For particular values of \( R \), he also gave some exact expressions and several identities. Coxeter (1935) has established further properties of \( \Phi_4^0 \) when the correlation coefficients \( \rho_{13}, \rho_{14} \) and \( \rho_{24} \) are all zero. When all the correlation coefficients are equal to \( \rho \), Ruben (1954) has shown that Schläfli's results yield a simple formula connecting \( \Phi_n^0 \) with \( \Phi_{n-2}^0 \); and he gives, in his Table 1, the values of \( \Phi_n^0 \) for

\[
1/\rho = 2, 3, \ldots, 12
\]

and

\[
n = 0, 1, 2, \ldots, 51 - 1/\rho.
\]
Another general approach to the problem is by means of the tetrachoric series, the details of which when \( n \) is 2 are given by Kendall (1945). This series has been generalized by Kendall (1941) and Moran (1948), who state explicit formulae for \( \Phi_n \) when \( n \) is 3 and 4; but although the tetrachoric series will always converge, it does so very slowly when the absolute values of the correlation coefficients are near unity, and some alternative approach then appears desirable, as David (1953) has indicated.

A direct attack is occasionally rewarding. Thus, when

\[
c_{pq} = \begin{cases} 
2, & p = q, \\
-1, & p - q = \pm 1, \\
0, & \text{otherwise},
\end{cases}
\]

Anis & Lloyd (1953) prove that \( \Phi_0^0 \) is \( 1/(n+1) \).

In what follows, we obtain a reduction formula for \( \Phi_n \), which permits us to express \( \Phi_3 \) and \( \Phi_4 \) as finite sums of single integrals of tabulated functions. These integrals must in general be evaluated by numerical quadrature, but under certain conditions we shall derive simple approximations to them. Schläfli's differential recurrence formula can be deduced from ours by putting

\[
a_1 = a_2 = \ldots = a_n = 0,
\]

and interpreting the result geometrically.
2. Our starting point is the system of partial differential equations represented by

\[ \frac{\partial \phi_n}{\partial \rho_{ij}} = \frac{\partial^2 \phi_n}{\partial x_i \partial x_j}. \]  

(3)

The equation for \( n \) equal to 2 has been known for some time. To establish (3) for all \( n \), we write \( \phi_n \) as the transform of its characteristic function:

\[ \phi_n = (2\pi)^{-n} \int \int \ldots \int \exp \left( -it'x - \frac{1}{2}t'Rt \right) dt_1 dt_2 \ldots dt_n. \]

Both sides of (3) then become

\[ -(2\pi)^{-n} \int \int \ldots \int t_i t_j \exp \left( -it'x - \frac{1}{2}t'Rt \right) dt_1 dt_2 \ldots dt_n. \]

We next differentiate \( \Phi_n \) with respect to all \( \{\rho_{ij}\} \):

\[ \frac{\partial \Phi_n}{\partial \rho_{12}} = \int_{a_1}^\infty \int_{a_1}^\infty \ldots \int_{a_n}^\infty (\frac{\partial \phi_n}{\partial \rho_{12}}) \, dx_1 \, dx_2 \ldots dx_n \]

\[ = \int_{a_1}^\infty \int_{a_1}^\infty \ldots \int_{a_n}^\infty \left( \int_{a_1}^\infty \int_{a_2}^\infty (\frac{\partial^2 \phi_n}{\partial x_1 \partial x_2}) \, dx_1 \, dx_2 \right) \, dx_3 \, dx_4 \ldots dx_n \]

\[ = \int_{a_1}^\infty \int_{a_1}^\infty \ldots \int_{a_n}^\infty \phi_n(a_1, a_2, x_3, x_4, \ldots, x_n; c_{11}, c_{12}, \ldots, c_{nn}) \, dx_3 \, dx_4 \ldots dx_n. \]  

(4)
To bring this into a more convenient form, we note that the conditional distributions of the normal multivariate distribution are expressible in terms of deviations from regression. In fact, let

\[ x'_1 = (x_1, x_2, \ldots, x_m), \]
\[ x'_2 = (x_{m+1}, x_{m+2}, \ldots, x_n), \]

and

\[ C = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}, \]

where \( C_{11} \) is \( m \times m \). If the inverse of \( C \), correspondingly partitioned, is

\[ R = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}, \]

then

\[ R^{-1}_{11} = C_{11} - C_{12} C_{22}^{-1} C_{21}, \]
\[ |R_{11}| = |C_{22}|/|C|, \]

and

\[ R_{21} R^{-1}_{11} = -C_{22}^{-1} C_{21}. \]

Partitioning \( x \) and \( C \) in (1), and using the last three equations, we get

\[ \phi_n = (2\pi)^{-\frac{m}{2}} |R_{11}|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} x'_1 R^{-1}_{11} x_1 \right\} \]
\[ \times (2\pi)^{-\frac{(n-m)}{2}} |C_{22}|^{\frac{1}{2}} \exp \left\{ -\frac{1}{2} (x_2 - R_{21} R^{-1}_{11} x_1)' C_{22} (x_2 - R_{21} R^{-1}_{11} x_1) \right\}. \]

(5)
The special case when \( m \) is 2 is relevant here. Put
\[
b_r = 
\frac{((\rho_{1r} - \rho_{2r} \rho_{12}) a_1 + (\rho_{2r} - \rho_{1r} \rho_{12}) a_2)}{(1 - \rho_{12}^2)} \quad (r = 3, 4, \ldots, n).
\]
Equation (4) becomes
\[
\partial \Phi_n / \partial \rho_{12} = (2\pi)^{-1} (1 - \rho_{12}^2)^{-\frac{3}{2}} \exp \left\{ -\frac{1}{2} (a_1^2 - 2\rho_{12} a_1 a_2 + a_2^2) / (1 - \rho_{12}^2) \right\}
\times \Phi_{n-2}(a_3 - b_3, a_4 - b_4, \ldots, a_n - b_n; c_{33}, c_{34}, \ldots, c_{nn}).
\]
(6)
Similar expressions for the other first-order partial derivatives can be written down at once.
3. For convenience in terminology, we consider any variance matrix \( R \) as defining a point \( P \) with co-ordinates \( \{\rho_{ij}\} \) in a space of \( \binom{n}{2} \) dimensions. When \( a_1, a_2, \ldots, a_n \) are fixed, the value of \( \Phi_n \) depends solely on the position of \( P \) and will be denoted by \( \Phi_n(P) \). Suppose that \( K \), with co-ordinates \( \{\kappa_{ij}\} \), is any point where the value of \( \Phi_n \) is obvious or can be calculated directly from the tables of \( \Phi_1 \) and \( \Phi_2 \). Let \( \{\lambda_{ij}(t)\} \) be the co-ordinates of the point \( L \) which divides \( KP \) in the ratio \( t:1-t \). Then

\[
\lambda_{ij}(t) = t\rho_{ij} + (1-t)\kappa_{ij}.
\]

Write \( R(t) \) for the matrix with unit diagonal elements and non-diagonal elements \( \{\lambda_{ij}(t)\} \); thus \( R(0) \) corresponds to the point \( K \) and \( R(1) \) to \( P \). Since

\[
R(t) = tR(1) + (1-t)R(0),
\]

it is positive definite for \( 0 < t \leq 1 \). Denote the inverse of \( R(t) \) by \( C(t) \), with elements \( c_{pq}(t) \). \( C(t) \) is positive definite for the same range of \( t \), and therefore the type of argument leading to equation (6) can be applied at any intermediate point on \( KP \), so that \( \partial \Phi_n / \partial \lambda_{12} \) at the point \( L \) is obtained from (6) on replacing \( \rho_{ij} \) by \( \lambda_{ij}(t) \) and \( c_{pq} \) by \( c_{pq}(t) \).

To derive the basic reduction formula, we integrate the differential element \( d\Phi_n \) along the line \( KP \), and obtain

\[
\Phi_n(P) = \Phi_n(K) + \sum_{i<j} \int_{\kappa_{ij}}^{\rho_{ij}} \frac{\partial \Phi_n}{\partial \lambda_{ij}}(L) \, d\lambda_{ij}. \tag{7}
\]

On coming to the numerical evaluation of (7), we shall change the variable of integration by writing

\[
\lambda_{ij} = \cos \theta_{ij}.
\]

This has the effect of removing \( (1 - \lambda_{ij}^2)^{-1} \) from the integrand, thus shortening the computations, and avoiding the singularity which occurs at the lower limit when \( \kappa_{ij} \) is \( \pm 1 \).
4. There is considerable flexibility in the choice of $K$, but two general rules can be suggested.

(i) Select a point near to $P$, although, as will appear when we give an example with four variables, the length of the path of integration in the space of the correlation coefficients is, by itself, no guide to the amount of care which we have to bestow on the process of numerical integration.

(ii) Choose $R(0)$ to be a matrix of rank $(n - 1)$ with all but one of its correlation coefficients the same as those of $R(1)$; the second term on the right of (7) then consists of a single integral. To prove that this is always possible, let $R_{11}$ be the first $(n - 1)$ rows and columns of $R$, and partition the remainder of $R$ as before. Denote by $R_{12}^*$ the vector obtained from $R_{12}$ on replacing $\rho_{n-1,n}$ by $\rho_{n-1,n}^*$. Since

$$|R| = |R_{11}|(1 - R_{21} R_{11}^{-1} R_{12}),$$

we have to solve

$$R_{21}^* R_{11}^{-1} R_{12}^* = 1$$

for $\rho_{n-1,n}^*$. On rearrangement,

$$(R_{21}^* - R_{21}) R_{11}^{-1} (R_{12}^* - R_{12}) + 2(R_{21}^* - R_{21}) R_{11}^{-1} R_{12} - (1 - R_{21} R_{11}^{-1} R_{12}) = 0,$$

which is a quadratic equation with positive discriminant. It therefore provides two real values of $\rho_{n-1,n}^*$ and either of these is the correlation coefficient of $x_{n-1}$ and

$$x_n^* = \psi_1 x_1 + \psi_2 x_2 + \ldots + \psi_{n-1} x_{n-1},$$

where $R_{11} \psi = R_{12}^*$. 
When \( n \) is 3,

\[
\rho_{23}^* = \rho_{12}\rho_{13} \pm \{(1 - \rho_{12}^2)(1 - \rho_{13}^2)\}^{1/2}
\]

and

\[
x_3^* = \{ (\rho_{13} - \rho_{12}\rho_{23}^*)x_1 + (\rho_{23}^* - \rho_{12}\rho_{13})x_2 \}/(1 - \rho_{12}^2).
\]

The tables of \( \Phi_1 \) and \( \Phi_2 \) suffice to calculate

\[
\Phi_3(K) = \Pr(x_1 > a_1, x_2 > a_2, x_3^* > a_3).
\]

When \( n \) is 4,

\[
(1 - \rho_{12}^2)\rho_{34}^* = \rho_{14}(\rho_{13} - \rho_{12}\rho_{23}) + \rho_{24}(\rho_{23} - \rho_{12}\rho_{13})
\]

\[
\pm \{(1 - \rho_{12}^2 - \rho_{13}^2 - \rho_{23}^2 + 2\rho_{12}\rho_{13}\rho_{23})(1 - \rho_{12}^2 - \rho_{14}^2 - \rho_{24}^2 + 2\rho_{12}\rho_{14}\rho_{24})\}^{1/2}.
\]

The coefficients in \( x_4^* \) are obtained by solving the system of equations

\[
\psi_1 + \rho_{12}\psi_2 + \rho_{13}\psi_3 = \rho_{14},
\]

\[
\rho_{12}\psi_1 + \psi_2 + \rho_{23}\psi_3 = \rho_{24},
\]

and

\[
\rho_{13}\psi_1 + \rho_{23}\psi_2 + \psi_3 = \rho_{34}^*,
\]

and are checked by evaluating

\[
\rho_{14}\psi_1 + \rho_{24}\psi_2 + \rho_{34}^*\psi_3 = 1.
\]
The combination of probabilities forming $\Phi_4(K)$ can be derived mechanically as follows. In the space $S$ of points $(x_1, x_2, x_3)$ let $S_1$, $S_2$, $S_3$ and $T$ be the regions defined by $x_1 > a_1$, $x_2 > a_2$, $x_3 > a_3$ and $x_4^* > a_4$ respectively. There will always exist some null relation of the form

$$\bar{S}_1 S_2 \bar{S}_3 T = 0,$$

where a bar signifies the complementary set. On expanding

$$(S - S_1) S_2 (S - S_3) T = 0,$$

we get

$$S_1 S_2 S_3 T = S_1 S_2 T + S_2 S_3 T - S_2 T.$$

$\Phi_4(K)$ therefore involves at most four trivariate normal integrals.
5. Our methods lead to the following expression for the trivariate normal integral:

\[ \Phi_3(P) = \Pr(x_1 > a_1, x_2 > a_2, x_3^* > a_3) \]

\[ + (2\pi)^{-1} \int_{\cos^{-1} \rho_{13}}^{\cos^{-1} \rho_{13}^*} \exp \left\{ -\frac{a_2^2 + a_3^2 - 2a_2a_3 \cos \theta}{2 \sin^2 \theta} \right\} \]

\[ \times \Phi_1 \left\{ \frac{(a_1 - \rho_{12}a_2 - \rho_{13}a_3) + (\rho_{13}a_2 + \rho_{12}a_3) \cos \theta - a_1 \cos^2 \theta}{\sin \theta (1 - \rho_{12}^2 - \rho_{13}^2 + 2\rho_{12}\rho_{13} \cos \theta - \cos^2 \theta)\dot{1}}; 1 \right\} d\theta. \quad (11) \]
Suppose, for example, that

\[ a_1 = -1.2, \quad a_2 = -1.0, \quad a_3 = 0.5 \]

and

\[ R = \begin{bmatrix} 1 & 0.7 & 0.2 \\ 1 & -0.4 \\ 1 \end{bmatrix}. \]

Equation (8) gives

\[ \rho_{23}^* = 0.14 \pm 0.699714. \]

Taking the critical value to be \(-0.559714\), and using (9),

\[ x_3^* = 1.160392x_1 - 1.371989x_2. \]

The inequalities \(x_1 < -1.2\) and \(x_2 > -1.0\) together imply \(x_3^* < -0.020481\), and so the set \(S_1S_2T\) is null. Hence

\[ S_1S_2T = S_2T, \]

giving

\[ \Phi_3(K) = \Pr(x_2 > -1.0, x_3^* > 0.5; \rho_{23}^* = -0.559714) = 0.587191. \]

Since \(\rho_{23}\) and \(\rho_{23}^*\) are negative here, we replace \(\theta\) by \(\pi - \theta\) in (11). The second stage of the work then consists in the evaluation of

\[
(2\pi)^{-1}\int_{\cos^{-1}(0.559714)}^{\cos^{-1}(0.4)} \exp\left\{-\left(0.625 - 0.5 \cos \theta\right)/\sin^2 \theta\right\}
\times \Phi_1\left\{\frac{-0.6 - 0.15 \cos \theta + 1.2 \cos^2 \theta}{\sin \theta \left(0.47 - 0.28 \cos \theta - \cos^2 \theta\right)}; 1\right\} d\theta.
\]
To achieve six significant figures in \( \Phi_3(P) \), we carry six as far as the final product for the integrand, where we step down to five, thus minimizing rounding-off errors. The integrand has a maximum near the lower limit and the contribution from this section was estimated by the three-eighths rule, using strips of \( \frac{1}{40} \) th of the interval. Thereafter, the integrand decreases steadily, but the first differences have a maximum, so that Gregory's formula was applied twice, firstly over nine strips of \( \frac{1}{40} \) th, and secondly over seven strips of \( \frac{1}{10} \) th. The final result is 0·016343, whence

\[
\Phi_3(P) = 0·603534.
\]

Of course, less work is required if less accuracy will suffice. In our example, \( \Phi_3(P) \) is given with an error of only one unit in the fourth decimal place merely by applying the trapezoidal rule to the integral.
6. When \( a_1, a_2, \ldots \) are all zero, equation (7) simplifies considerably, and we are easily led to the known results for \( n \) equal to 2 and 3. Taking \( K \) at the origin,

\[
\Phi^0_2(P) = \frac{1}{2} + (2\pi)^{-1} \int_{\cos^{-1} \rho_{12}}^{\pi} d\theta = \{\cos^{-1}(-\rho_{12})\}/2\pi;
\]

and

\[
\Phi^0_3(P) = \frac{1}{8} + (4\pi)^{-1} \left( \int_{\cos^{-1} \rho_{12}}^{\pi} d\theta + \int_{\cos^{-1} \rho_{13}}^{\pi} d\theta + \int_{\cos^{-1} \rho_{23}}^{\pi} d\theta \right) = \{\cos^{-1}(-\rho_{12}) + \cos^{-1}(-\rho_{13}) + \cos^{-1}(-\rho_{23}) - \pi\}/4\pi.
\]

When \( n \) is 4, put

\[
\sigma_{pq}(t) = c_{pq}(t)/\{c_{pp}(t)c_{qq}(t)\}^{\frac{1}{4}},
\]

and (7) becomes

\[
\Phi^0_4(P) = \Phi^0_4(K) + (1/4\pi^2) \sum_{p < q} \rho_{pq} \cos^{-1} \sigma_{pq}(1 - \lambda^2_{ij})^{-\frac{1}{4}} d\lambda_{ij},
\]

(12)

where \((p, q, i, j)\) is a permutation of \((1, 2, 3, 4)\).
Two approximations follow. If all the \{\rho_{ij}\} are small, we take \(K\) at the origin and get
\[
\sigma_{ij} \simeq -t \rho_{ij}.
\]
The integrals in (12) now group themselves in pairs to give
\[
\Phi_4^0(P) \simeq \left\{ \cos^{-1}(-\rho_{12}) \cos^{-1}(-\rho_{34}) + \cos^{-1}(-\rho_{13}) \cos^{-1}(-\rho_{24}) \\
+ \cos^{-1}(-\rho_{14}) \cos^{-1}(-\rho_{23}) \right\}/4\pi^2 - \frac{1}{8}. \tag{13}
\]
If we expand (13), ignoring quantities of order \(\rho^3\), we get the first three terms of Moran’s formula (3).

When all the differences \{\rho_{ij} - \kappa_{ij}\} are small, and provided that \(\sigma_{pq}(0)\) is not close to \(\pm 1\), we can write
\[
\cos^{-1} \sigma_{pq}(t) \simeq t \cos^{-1} \sigma_{pq}(1) + (1 - t) \cos^{-1} \sigma_{pq}(0) = \alpha_{pq} + \beta_{pq} \lambda_{ij}(t).
\]
Under these conditions
\[
\int_{\kappa_{ij}}^{\rho_{ij}} \cos^{-1} \sigma_{pq}(1 - \lambda_{ij}^2)^{-\frac{1}{2}} d\lambda_{ij} \simeq \alpha_{pq}(\cos^{-1} \kappa_{ij} - \cos^{-1} \rho_{ij}) + \beta_{pq}\{(1 - \kappa_{ij}^2)^{\frac{1}{2}} - (1 - \rho_{ij}^2)^{\frac{1}{2}}\}. \tag{14}
\]
This approximation cannot be used when \(R(0)\) is a matrix of rank 3, for the equation
\[
(1 - \sigma_{pq}^2) = (1 - \lambda_{ij}^2) |R(t)| \left| \begin{array}{c} 1 \\
1 \\
1 \\
1
\end{array} \right| \begin{array}{ccc} \lambda_{pt} & \lambda_{pj} & 1 \\
1 & \lambda_{ij} & 1 \\
1 & \lambda_{ij} & 1 \end{array}
\]
then implies that
\[
\sigma_{pq}(0) = \pm 1.
\]

A brief table of \(\Phi_4^0\) is given in an appendix, in order to extend the range of (14) by providing additional starting-points \(K\).
7. We shall illustrate the calculation of $\Phi_4^0$ by an example in which all the correlation coefficients are fairly large in absolute magnitude. Suppose that

$$R = \begin{bmatrix}
1 & -0.60 & 0.85 & 0.75 \\
1 & -0.70 & -0.80 \\
1 & 0.65 \\
1 & & & &
\end{bmatrix}.$$ 

Equation (10) gives

$$0.64 \rho_{34}^* = 0.4745 \pm 0.14928580.$$ 

Taking the positive sign

$$\rho_{34}^* = 0.97466531$$

and

$$x_4^* = -0.287x_1 - 0.234x_2 + 1.055x_3.$$ 

Since $S_1 S_2 \bar{S}_3 T$ is null,

$$\Phi_4^0(K) = \Pr(x_1 > 0, x_2 > 0, x_4^* > 0; \rho_{12} = -0.60, \rho_{14} = 0.75, \rho_{24} = -0.80)$$

$$= 0.067487.$$
Turning to the second term in (12),

\[
\sigma_{12} = -
\begin{vmatrix}
-0.60 & 0.85 & 0.75 & 1 & -0.70 & -0.80 & 1 & 0.85 & 0.75 \\
-0.70 & 1 & \cos \theta & -0.70 & 1 & \cos \theta & 0.85 & 1 & \cos \theta \\
-0.80 & \cos \theta & 1 & -0.80 & \cos \theta & 1 & 0.75 & \cos \theta & 1
\end{vmatrix}^{-\frac{1}{2}},
\]

and we thus have the problem of computing

\[
-(1/4\pi^2)\int_{\cos^{-1}(0.65)}^{\cos^{-1}(0.97...)} \cos^{-1}\left\{ \frac{-0.595 + 1.205 \cos \theta - 0.6 \cos^2 \theta}{(-0.13 + 1.12 \cos \theta - \cos^2 \theta)^{\frac{1}{2}} (-0.285 + 1.275 \cos \theta - \cos^2 \theta)^{\frac{1}{2}}} \right\} d\theta.
\]

This integral presents two features which make the computations more laborious than in the previous example. In the first place, \(\cos \theta\) must be recorded to eight decimal places in order to obtain six in the argument of the inverse cosine. Secondly, this argument is 1 at the lower limit, and because

\[
\cos^{-1}(1-x) = \sqrt{(2x)(1+x/12+3x^2/160+5x^3/896+...)}
\]

the integrand behaves like \(x^\frac{1}{4}\) when \(x\) is close to zero, so that the usual methods of numerical integration break down. The difficulty is overcome by using a formula given by Jeffreys & Jeffreys (1946), §9·092, namely

\[
\int_0^{3^h} (\alpha x^\frac{1}{4} + \beta x^\frac{3}{4} + \gamma x^\frac{7}{4}) dx = h(6 \sqrt{3} y_1/7 + 12 \sqrt{6} y_2/35 + 16 y_3/35).
\]

This was applied with strips of \(\frac{1}{250}\)th of the interval, and followed by Gregory's formula over seven strips of \(\frac{1}{250}\)th, eight strips of \(\frac{1}{50}\)th and eight strips of \(\frac{1}{10}\)th. The final result is 0·025164, whence

\[
\Phi_4'(P) = 0.042323.
\]
To test the adequacy of (14), we now work through the same example taking

\[
R(t) = \begin{bmatrix}
1 & -1 + 0.40t & 1 - 0.15t & 1 - 0.25t \\
1 & -1 + 0.30t & -1 + 0.20t \\
1 & 1 - 0.35t \\
1 & & & 
\end{bmatrix}
\]

\(R(0)\) is a matrix of rank 1. It corresponds to the system of relations

\[x_1 = -x_2 = x_3 = x_4,
\]

from which we infer that

\[
\Phi_4^0(K) = 0.
\]

The quantities \(\{\sigma_{pq}(t)\}\) are computed from the adjoint of \(R(t)\), namely,

\[
R(t) = t^2 \begin{bmatrix}
0.2175 - 0.042t & -0.0775 + 0.01225t & -0.1675 + 0.037t & -0.1275 + 0.0285t \\
0.1475 - 0.02625t & 0.1025 - 0.02375t & 0.1225 - 0.027775t \\
0.1975 - 0.04t & 0.0725 - 0.014t \\
0.1775 - 0.036t & & & 
\end{bmatrix}
\]

As each row of \(R(t)\) is completed, it is checked by forming its product—theoretically zero—with some other row of \(R(t)\). Of the six integrals to be evaluated, a typical one is

\[
I_{14} = -0.25 \int_0^1 \left\{ 1 - (1 - 0.25t)^2 \right\}^{0.4} \cos^{-1} \left( \frac{0.1025 - 0.02375t}{(0.1475 - 0.02625t)^{0.4} (0.1975 - 0.04t)^{0.4}} \right) dt
\]

\[
= -\int_0^{\cos^{-1}(0.75)} \cos^{-1} \left( \frac{0.0075 + 0.095 \cos \theta}{(0.0425 + 0.105 \cos \theta)^{0.4} (0.0375 + 0.160 \cos \theta)^{0.4}} \right) d\theta.
\]

The values taken by the inverse cosine at the limits of integration are 0.92662 and 0.96446; and

\[0.92662 (1 - t) + 0.96446t = 1.07798 - 0.15136(1 - 0.25t).
\]
Hence, using (14),

\[ I_{14} \approx -1.07798 \cos^{-1}(0.75) + 0.15136 (1 - 0.75^2)^{1/2} = -0.67898. \]

On the other hand, applying Gregory with five strips to the second formulation,

\[ I_{14} = -0.67827. \]

Although the length of the path of integration in the space of the \( \{\rho_{ij}\} \) is double what it was previously—0.705 against 0.325—and although the number of integrations has been increased from one to six, yet the total amount of work in evaluating \( \Phi_4^0(P) \) to six decimal places is roughly the same as before. The exact and approximate results for all six integrals are collected in the table below. Approximation (14) therefore leads to an estimate of \( \Phi_4^0(P) \) differing from the exact value by less than one unit in the fourth decimal place; and it thus appears to work quite satisfactorily, considering that the differences \( \{\rho_{ij} - \kappa_{ij}\} \) are not particularly small.
<table>
<thead>
<tr>
<th>Integral</th>
<th>Approximation (14)</th>
<th>Numerical integration</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I_{12}$</td>
<td>1.08628</td>
<td>1.08643</td>
</tr>
<tr>
<td>$I_{13}$</td>
<td>-0.40438</td>
<td>-0.40371</td>
</tr>
<tr>
<td>$I_{14}$</td>
<td>-0.67898</td>
<td>-0.67827</td>
</tr>
<tr>
<td>$I_{23}$</td>
<td>1.80380</td>
<td>1.80439</td>
</tr>
<tr>
<td>$I_{24}$</td>
<td>1.60801</td>
<td>1.60864</td>
</tr>
<tr>
<td>$I_{34}$</td>
<td>-1.74700</td>
<td>-1.74661</td>
</tr>
<tr>
<td>Total</td>
<td>1.66773</td>
<td>1.67087</td>
</tr>
<tr>
<td>$\div 4\pi^2$</td>
<td>0.042244</td>
<td>0.042324</td>
</tr>
</tbody>
</table>
8. The calculation of $\Phi_4$ by our method is similar to, but more laborious than, the calculation of $\Phi_3$. Once again, single integrals will suffice. If suggestion (ii) is accepted, $\Phi_4(K)$ can be expressed as a trivariate normal integral and is evaluated as described above; the integrand of the remaining term is a product of probabilities and ordinates of the normal bivariate distribution. For $\Phi_5$ and $\Phi_6$, double integrals are required, and, in general, the effect of repeatedly applying equation (7) is to reduce the multiplicity of a normal integral by more than half. Whether our results offer a practical method of approach when $n$ exceeds 4 is a moot point, but they are perhaps worth considering when values of $\Phi_5$, $\Phi_6$, ... are urgently needed.

I am indebted to a referee for simplifying my original proof of equation (3).
APPENDIX

Our method of calculating $\Phi_n$ depends on knowing the values of this function for certain values of $R$. Such knowledge need not be systematic, however—as would be essential if we were interpolating—and consequently the values of $\Phi_4^0$ given below, although forming a rather haphazard collection, may nevertheless be useful. Since the results of Anis & Lloyd (1953) and Ruben (1954) are available in recent issues of this journal, they are not reproduced here.

(A) \[
R = \begin{bmatrix}
1 & \rho_{12} & 0 & 0 \\
1 & \rho_{23} & 0 \\
1 & \rho_{34} \\
1
\end{bmatrix}.
\]

Put \[\rho_{12} = -\cos \alpha, \quad \rho_{23} = -\cos \beta, \quad \rho_{34} = -\cos \gamma,\]
where \[0 \leq \alpha \leq \frac{1}{2}\pi, \quad 0 \leq \beta \leq \pi, \quad 0 \leq \gamma \leq \frac{1}{2}\pi.\]

The relationship between the function $f(\alpha, \beta, \gamma)$ of Schläfli, the infinite sum $S(\alpha, \beta, \gamma)$ of Coxeter, and $\Phi_4^0$, is expressed by

\[\Phi_4^0 = f(\alpha, \beta, \gamma)/16 = S(\frac{1}{2}\pi - \alpha, \beta, \frac{1}{2}\pi - \gamma)/8\pi^2.\]
The following table was computed from a set of exact values of \(f(\alpha, \beta, \gamma)\) given by Schläfli (third paper, p. 98); the first ten are also derived by Coxeter:

<table>
<thead>
<tr>
<th>(\rho_{12})</th>
<th>(\rho_{23})</th>
<th>(\rho_{34})</th>
<th>(\Phi^0_4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.500000</td>
<td>0.500000</td>
<td>0.500000</td>
<td>0.008333</td>
</tr>
<tr>
<td>0.707107</td>
<td>0.500000</td>
<td>0.500000</td>
<td>0.002604</td>
</tr>
<tr>
<td>0.500000</td>
<td>0.707107</td>
<td>0.500000</td>
<td>0.000868</td>
</tr>
<tr>
<td>0.809017</td>
<td>0.500000</td>
<td>0.500000</td>
<td>0.000069</td>
</tr>
<tr>
<td>0.309017</td>
<td>0.500000</td>
<td>0.309017</td>
<td>0.013264</td>
</tr>
<tr>
<td>0.809017</td>
<td>0.809017</td>
<td>0.500000</td>
<td>0.001389</td>
</tr>
<tr>
<td>0.309017</td>
<td>0.809017</td>
<td>0.809017</td>
<td>0.000278</td>
</tr>
<tr>
<td>0.500000</td>
<td>0.309017</td>
<td>0.809017</td>
<td>0.005278</td>
</tr>
<tr>
<td>0.809017</td>
<td>0.309017</td>
<td>0.309017</td>
<td>0.000417</td>
</tr>
<tr>
<td>0.309017</td>
<td>0.809017</td>
<td>0.809017</td>
<td>0.004583</td>
</tr>
<tr>
<td>0.500000</td>
<td>0.500000</td>
<td>0.790569</td>
<td>0.000408</td>
</tr>
<tr>
<td>0.500000</td>
<td>0.790569</td>
<td>0.250000</td>
<td>0.001631</td>
</tr>
<tr>
<td>0.790569</td>
<td>0.250000</td>
<td>0.790569</td>
<td>0.002446</td>
</tr>
<tr>
<td>0.500000</td>
<td>0.309017</td>
<td>0.925615</td>
<td>0.000217</td>
</tr>
<tr>
<td>0.309017</td>
<td>0.925615</td>
<td>0.135045</td>
<td>0.000930</td>
</tr>
<tr>
<td>0.925615</td>
<td>0.135045</td>
<td>0.809017</td>
<td>0.001624</td>
</tr>
<tr>
<td>0.135045</td>
<td>0.809017</td>
<td>0.500000</td>
<td>0.002301</td>
</tr>
</tbody>
</table>
From each value of $\Phi_4^0$, five others can be generated by changing the signs of the variables $x_1, x_2, x_3$ and $x_4$. For example,

$$
\Phi_4^0 \begin{bmatrix}
1 & \frac{1}{2} & 0 & 0 \\
1 & -\frac{1}{2} & 0 \\
1 & -\frac{1}{2} \\
1 & 1 \\
\end{bmatrix}
= \Phi_3^0 \begin{bmatrix}
1 & -\frac{1}{2} & 0 \\
1 & 1 & -\frac{1}{2} \\
1 \\
\end{bmatrix}
- \Phi_4^0 \begin{bmatrix}
1 & -\frac{1}{2} & 0 & 0 \\
1 & 1 & -\frac{1}{2} \\
1 \\
\end{bmatrix}.
$$

Thus, the set of six associated with the first tabular entry is as follows:

<table>
<thead>
<tr>
<th>Formation</th>
<th>$\rho_{12}$</th>
<th>$\rho_{23}$</th>
<th>$\rho_{34}$</th>
<th>$\Phi_4^0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$-x_1$</td>
<td>$-\frac{1}{2}$</td>
<td>$-\frac{1}{2}$</td>
<td>$-\frac{1}{2}$</td>
<td>$\frac{1}{6}$</td>
</tr>
<tr>
<td>$-x_1, -x_2$</td>
<td>$-\frac{1}{2}$</td>
<td>$-\frac{1}{2}$</td>
<td>$-\frac{1}{2}$</td>
<td>$\frac{1}{6}$</td>
</tr>
<tr>
<td>$-x_2$</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{1}{6}$</td>
</tr>
<tr>
<td>$-x_1, -x_4$</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{1}{6}$</td>
</tr>
<tr>
<td>$-x_1, -x_3$</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{1}{6}$</td>
</tr>
</tbody>
</table>
(B) Two more values of $\Phi_4^0$ are obtainable by elementary combinatorial methods, based on a set of uncorrelated random variables $y_1, y_2, ..., y_5$, all of which have the same distribution and variance $\frac{1}{2}$. For example, the first $\mathbf{R}$ is the variance matrix of $y_3 - y_1, y_4 - y_1, y_3 - y_2, y_4 - y_2$; and, out of $4!$ ordered arrangements of $y_1, y_2, y_3, y_4$, there are $(2!)^2$ which ensure that all four differences are positive. Hence $\Phi_4^0$ is $\frac{1}{6}$ whether or not anything further is specified about the distribution of $x_1, x_2, x_3$ and $x_4$.

<table>
<thead>
<tr>
<th>Formation</th>
<th>$\mathbf{R}$</th>
<th>$\Phi_4^0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>1 $\frac{1}{2}$ $\frac{1}{2}$ 0</td>
<td>$\frac{1}{6}$</td>
</tr>
<tr>
<td>14</td>
<td>1 $\frac{1}{2}$ 0 $\frac{1}{2}$</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>1 $\frac{1}{2}$ $\frac{1}{2}$ 1</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>1 $\frac{1}{2}$ 0 $\frac{1}{2}$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Formation</th>
<th>$\mathbf{R}$</th>
<th>$\Phi_4^0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>1 $\frac{1}{2}$ $\frac{1}{2}$ 0</td>
<td>$\frac{3}{5}$</td>
</tr>
<tr>
<td>13</td>
<td>1 $\frac{1}{2}$ $\frac{1}{2}$ 0</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>1 $\frac{1}{2}$ $\frac{1}{2}$ 1</td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>1 $\frac{1}{2}$ $\frac{1}{2}$ 1</td>
<td></td>
</tr>
</tbody>
</table>

The first $\mathbf{R}$ belongs to a set of four; and the second, to a set of six.
REFERENCES


Schläfli, L. (1858). On the multiple integral $\int^n dx \, dy \, \ldots \, dz$ whose limits are

$$p_1 = a_1 x + b_1 y + \ldots + h_1 z > 0, \quad p_2 > 0, \ldots, p_n > 0 \quad \text{and} \quad x^2 + y^2 + \ldots + z^2 < 1.$$  

MAXIMUM LIKELIHOOD ESTIMATION OF THE POLYCHORIC CORRELATION COEFFICIENT

ULF OLSSON

UNIVERSITY OF UPPSALA

The polychoric correlation is discussed as a generalization of the tetrachoric correlation coefficient to more than two classes. Two estimation methods are discussed: Maximum likelihood estimation, and what may be called "two-step maximum likelihood" estimation. For the latter method, the thresholds are estimated in the first step. For both methods, asymptotic covariance matrices for estimates are derived, and the methods are illustrated and compared with artificial and real data.

Key words: ordinal data, polychoric correlation.
1. Introduction

Especially in the behavioral sciences, data are often recorded as ordinal variables with only a few scale steps. Examples of such variables are attitude items, rating scales, Likert items and the like. Typical cases are when a subject is asked to report some attitude on scales like

(a) approve don’t know disapprove or
(b) approve strongly approve don’t know disapprove disapprove strongly.

When analyzing this kind of data, a common approach is to assign integer values to each category [for example 1, 2 and 3 in Example (a) and 1 through 5 in Example (b)] and proceed in the analysis as if the data had been measured on an interval scale with desired distributional properties. To quote Wainer and Thissen [1976], in such cases “a quick and easy approach is to assume normality and be on your merry way”.
Although many statistical methods seem to be fairly robust against this kind of deviation from the distributional assumptions—at least in not-so-extreme cases—there are instances when this approach may lead to erroneous results.

For example, Olsson [1979] showed that application of factor analysis to discrete data may lead to incorrect conclusions regarding the number of factors, and to biased estimates of the factor loadings, especially when the distributions of the observed variables are skewed in opposite directions. This is mainly due to biased estimates of the correlations. Thus, there seems to be some need for correlation estimates which are more viable when the observed data are ordinal with only a few scale steps.
In this paper, we shall discuss the maximum likelihood estimation of correlation coefficients from ordinal data. In short, our problem may be summarized as follows: We observe two ordinal variables, \( x \) and \( y \). These are classified into \( s \) and \( r \) categories, respectively. A cross-tabulation of \( x \) by \( y \) gives the observed frequencies as denoted in Table 1. We further assume that underlying \( x \) and \( y \) there are some latent variables, \( \xi \) and \( \eta \), which are bivariate normally distributed. The relation between \( x \) and \( \xi \) may be written

\[
x = 1 \quad \text{if } \xi < a_1, \\
x = 2 \quad \text{if } a_1 \leq \xi < a_2, \\
x = 3 \quad \text{if } a_2 \leq \xi < a_3, \\
\vdots \\
x = s \quad \text{if } a_{s-1} \leq \xi,
\]

and correspondingly for \( y \). The parameters \( a_i \) are usually termed thresholds. The problem is to estimate the correlation \( \rho \) between \( \xi \) and \( \eta \), given data in the form of Table 1. This is a generalization of the arguments behind the familiar tetrachoric correlation coefficient to polytomous variables.
TABLE 1

The General Form of the Raw Data: a Cross-tabulation of $x$ by $y$.

<table>
<thead>
<tr>
<th>$y$</th>
<th>$b_1$</th>
<th>$b_2$</th>
<th>...</th>
<th>$b_{r-1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x$</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>...</td>
</tr>
<tr>
<td>1</td>
<td>$n_{11}$</td>
<td>$n_{12}$</td>
<td>$n_{13}$</td>
<td>...</td>
</tr>
<tr>
<td>$a_1$</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$a_2$</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$a_{s-1}$</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$s$</td>
<td>$n_{s1}$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$a_i$ and $b_j$ denote thresholds, where $a_0 = b_0 = -\infty$ and $a_s = b_r = +\infty$. 
We shall discuss estimation of $\rho$ by means of the maximum likelihood method. Even given the method of estimation, the problem may be solved in at least two different ways. One way is to estimate $\rho$ and the thresholds simultaneously. Alternatively, the thresholds are first estimated as the inverse of the normal distribution function, evaluated at the cumulative marginal proportions of the table, and the maximum likelihood estimate of $\rho$ is then computed given the thresholds. This may be called a "two-step maximum likelihood" procedure. The latter approach has the advantage of greater ease in the numerical computations, although the former is formally more correct. One of the points discussed in the paper is the differences in results from the two methods in some reasonable cases.

In Section 2 we summarize the results of some earlier writers in the area. In Section 3 the likelihood equations are derived, and in Section 4 we derive asymptotic standard errors of the estimates. Section 5 contains a numerical computer study, based on both true data and on Monte Carlo simulations. There we also analyze some real data. Finally our results are summarized in Section 6.
2. Earlier Research

2 × 2 Tables: The Tetrachoric Correlation

The tetrachoric correlation coefficient was suggested by Pearson [1901] as a measure of bivariate normal correlation, when only data from a 2 × 2 cross-classification of the data are available. Pearson also supplied formulae for the standard error of the correlation.

The tetrachoric correlation may be obtained by solving $\rho$ from

\[
\Phi(h, k; \rho) = \left[2\pi(1 - \rho^2)^{1/2}\right]^{-1} \int_{-\infty}^{h} \int_{-\infty}^{k} \exp \left[ -\frac{x^2 - 2\rho xy + y^2}{2(1 - \rho^2)} \right] dx \, dy. \tag{1}
\]

The method suggested by Pearson [1901] was to expand the right-hand side of (1) into a series expansion in $\rho$, the so called tetrachoric series. This leads to a polynomial equation in $\rho$, where the degree of the polynomial depends on the number of terms in the series expansion.
Hamdan [1970] showed that the tetrachoric $r$ is equivalent to the maximum likelihood estimate of $\rho$ from a 2 by 2 table. This is not unexpected, since the estimation problem involves three parameters ($\rho$, $h$ and $k$ where $h$ and $k$ are the thresholds), and the data supplies three independent statistics in the table, for example $n_{1.}$, $n_{.1}$ and $n_{11}$. Thus, the estimation problem is just identified. The advantage of the maximum likelihood approach is that asymptotic standard errors are easily obtained through the inverse of the second order derivative of the log likelihood.

The tetrachoric correlation is a biased estimate of the corresponding true correlation. Brown and Bendetti [1977] showed that the expected cell frequencies are critical for the bias. If no expected frequencies are less than 5, the bias is negligible. This applies also for the bias of the standard error if $\rho = 0$. If $\rho \neq 0$, Brown and Bendetti recommended the standard error based on ML-estimation, but even this converges fairly slowly towards true $\sigma_\rho$ with increasing $n$. It seems that expected frequencies of at least 10 are necessary in order to obtain reasonable estimates of the standard error.
Computer routines for the numerical solution of (1) are now available as packed programs at many computer installations. Froemel [1971] made a comparison of some routines. The best routine to date is probably that by Kirk [1973], who attacks the problem by Gaussian quadrature and Newton-Raphson iteration.
$r \times s$ Tables: The Polychoric Correlation

Several coefficients have been suggested as measures of association in contingency tables; we shall here confine ourselves to those which, like the tetrachoric correlation, use an assumption of an underlying bivariate normal distribution, for which the correlation is to be estimated.

Pearson [Note 1] suggested that the mean square contingency, which is based on the usual $\chi^2$, could be used to estimate $\rho$ from a polychoric table. Ritchie-Scott [1918] developed a coefficient based on a weighted mean of all possible tetrachoric correlations which may be computed from the table. Pearson and Pearson [1922], influenced by the result of Ritchie-Scott, suggested simpler methods for larger tables.
Lancaster and Hamdan [1964] showed that the mean square contingency does not work well when the classification is crude. Instead, using the theory of orthonormal functions, they generalized the tetrachoric series expansion to the polychoric case. Pearson's corrected $\phi^2$ was equated to a series which they called the polychoric series and the corresponding polynomial equation was solved for $\rho$. The method does not produce standard errors of the estimates; confidence intervals are instead obtained by inserting 2.5 and 97.5 percent limits for $\chi^2$ in $\phi^2$ and solving the equations. Hamdan [1968] showed that Pearson's [1901] tetrachoric series is a special case of the polychoric one. Hamdan [1971] gave computing formulae for smaller tables, and Martinson and Hamdan [1975] designed a computer program for the method. Martinson and Hamdan [1971] used a Maximum Likelihood approach to estimate $\rho$ given the thresholds, and did also present formulae for the asymptotic standard error. The latter formulae, however, do not take into account that the thresholds are estimated from the data, and that they are thus also subject to random errors.
In all the above methods, the thresholds are regarded as fixed, although already Pearson and Pearson [1922] were aware that this might not be ideal. For the tetrachoric case this does not matter, since, as noted above, the estimation problem is just identified. For 3 by 3 tables, for example, we have 5 parameters ($\rho$ and two thresholds for each variable) but 8 independent proportions, i.e. the problem is over-identified. Tallis [1962] attacked the problem via maximum likelihood estimation of $\rho$ and thresholds simultaneously in $2 \times 2$ and $3 \times 3$ tables. For $2 \times 2$ tables the results should be the tetrachoric correlation [cf. Hamdan, 1970], but for $3 \times 3$ tables the results may differ from Lancaster and Hamdan’s [1964] polychoric correlation.
3. Derivation of the Likelihood Equations

The data consists of an array of observed frequencies \( n_{ij} : i = 1, 2, \ldots, s; j = 1, 2, \ldots, r \), as given in Table 1.

If we denote by \( \pi_{ij} \) the probability that an observation falls into cell \((i, j)\), the likelihood of the sample is

\[
L = C \cdot \prod_{i}^{s} \prod_{j}^{r} \pi_{ij}^{n_{ij}}
\]

where \( C \) is a constant. Taking logarithms,

\[
l = \ln L = \ln C + \sum_{i=1}^{s} \sum_{j=1}^{r} n_{ij} \ln \pi_{ij}.
\]

The thresholds for \( x \) are denoted by \( a_{i}, i = 0 \cdots s \) and the thresholds for \( y \) by \( b_{j}, j = 0 \cdots r \), where \( a_{0} = b_{0} = -\infty \) and \( a_{s} = b_{r} = +\infty \). It follows that

\[
\pi_{ij} = \Phi_{2}(a_{i}, b_{j}) - \Phi_{2}(a_{i-1}, b_{j}) - \Phi_{2}(a_{i}, b_{j-1}) + \Phi_{2}(a_{i-1}, b_{j-1})
\]

where \( \Phi_{2} \) is the bivariate normal distribution function with correlation \( \rho \).
Case 1: All Parameters Are Estimated Simultaneously

The parameters to be estimated are $\rho, a_1 \cdots a_{s-1}, b_1 \cdots b_{r-1}$. Partial differentiation of $l$ with respect to these parameters yields

\begin{align*}
\frac{\partial l}{\partial \rho} &= \sum_{i=1}^{s} \sum_{j=1}^{r} \frac{n_{ij}}{\pi_{ij}} \frac{\partial \pi_{ij}}{\partial \rho} \\
\frac{\partial l}{\partial a_k} &= \sum_{i=1}^{s} \sum_{j=1}^{r} \frac{n_{ij}}{\pi_{ij}} \frac{\partial \pi_{ij}}{\partial a_k} \\
\frac{\partial l}{\partial b_m} &= \sum_{i=1}^{s} \sum_{j=1}^{r} \frac{n_{ij}}{\pi_{ij}} \frac{\partial \pi_{ij}}{\partial b_m}.
\end{align*}

Since $\partial \Phi_2(u, v)/\partial \rho = \phi_2(u, v)$ where $\phi_2$ is the bivariate normal density function, (see Tallis, 1962, p 344; see also Johnson & Kotz, 1972, p 44), it follows that

\begin{equation}
\frac{\partial \pi_{ij}}{\partial \rho} = \phi_2(a_i, b_j) - \phi_2(a_{i-1}, b_j) - \phi_2(a_i, b_{j-1}) + \phi_2(a_{i-1}, b_{j-1}).
\end{equation}
Therefore, (5) may be written

\[
\frac{\partial I}{\partial \rho} = \sum_{i=1}^{s} \sum_{j=1}^{r} \frac{n_{ij}}{\pi_{ij}} \{\phi_2(a_i, b_j) - \phi_2(a_{i-1}, b_j) - \phi_2(a_i, b_{j-1}) + \phi_2(a_{i-1}, b_{j-1})\}. \tag{9}
\]

In (6), it is evident that

\[
\frac{\partial \pi_{ij}}{\partial a_k} = \begin{cases} 
0 & \text{if } i \neq k \text{ and } i \neq k + 1, \text{ i.e. if the formula for } \pi_{ij} \text{ does not contain } a_k \\
\frac{\partial \Phi_2(a_k, b_j)}{\partial a_k} - \frac{\partial \Phi_2(a_k, b_{j-1})}{\partial a_k} & \text{if } k = i \\
- \frac{\partial \Phi_2(a_k, b_j)}{\partial a_k} + \frac{\partial \Phi_2(a_k, b_{j-1})}{\partial a_k} & \text{if } k = i - 1.
\end{cases} \tag{10}
\]
Thus, in (6) it suffices to let $i$ go from $k$ to $k + 1$. Therefore, (6) may be written

\[
\frac{\partial l}{\partial a_k} = \sum_{j=1}^{r} \frac{n_{k,j}}{\pi_{k,j}} \left\{ \frac{\partial \Phi_2(a_k, b_j)}{\partial a_k} - \frac{\partial \Phi_2(a_k, b_{j-1})}{\partial a_k} \right\} \\
+ \frac{n_{k+1,j}}{\pi_{k+1,j}} \left\{ -\frac{\partial \Phi_2(a_k, b_j)}{\partial a_k} + \frac{\partial \Phi_2(a_k, b_{j-1})}{\partial a_k} \right\} \\
= \sum_{j=1}^{r} \left( \frac{n_{k,j}}{\pi_{k,j}} - \frac{n_{k+1,j}}{\pi_{k+1,j}} \right) \left\{ \frac{\partial \Phi_2(a_k, b_j)}{\partial a_k} - \frac{\partial \Phi_2(a_k, b_{j-1})}{\partial a_k} \right\}.
\]
Also, if we let \( \phi_1 \) and \( \Phi_1 \) denote univariate normal density and distribution function, respectively,

\[
\frac{\partial \Phi_2(u, v)}{\partial u} = \phi_1(u) \cdot \Phi_1 \left\{ \frac{(v - \rho u)}{(1 - \rho^2)^{1/2}} \right\}
\]

[Tallis, 1962, p 346]. Equation (6) may now be written as

\[
\frac{\partial l}{\partial a_k} = \sum_{j=1}^r \left( \frac{n_{kj}}{n_{k+1,j}} - \frac{n_{k+1,j}}{n_{k+1,j}} \right) \cdot \phi_1(a_k) \cdot \left[ \Phi_1 \left\{ \frac{(b_j - \rho a_k)}{(1 - \rho^2)^{1/2}} \right\} - \Phi_1 \left\{ \frac{(b_{j-1} - \rho a_k)}{(1 - \rho^2)^{1/2}} \right\} \right].
\]

From the symmetry it also follows that

\[
\frac{\partial l}{\partial b_m} = \sum_{i=1}^s \left( \frac{n_{im}}{n_{i,m+1}} - \frac{n_{i,m+1}}{n_{i,m+1}} \right) \cdot \phi_1(b_m) \cdot \left[ \Phi_1 \left\{ \frac{(a_i - \rho b_m)}{(1 - \rho^2)^{1/2}} \right\} - \Phi_1 \left\{ \frac{(a_{i-1} - \rho b_m)}{(1 - \rho^2)^{1/2}} \right\} \right].
\]

Equations (9), (13) and (14) constitute the set of first order derivatives of the log likelihood.
Case 2: The Thresholds are Computed from the Marginals

Most earlier researchers in this area [Pearson & Pearson, 1922; Lancaster & Hamdan, 1964; Hamdan, 1971; Martinson & Hamdan, 1971, 1975] have regarded the thresholds as given from the cumulative marginal proportions of the table. Although theoretically non-optimal, this approach has the advantage of reducing the computational labor.

In this case, the equation system to be solved is

\[
\frac{\partial l}{\partial \rho} = \sum_{i=1}^{s} \sum_{j=1}^{r} \frac{n_{ij}}{\pi_{ij}} [\phi_2(a_i, b_j) - \phi_2(a_{i-1}, b_j) - \phi_2(a_i, b_{j-1}) + \phi_2(a_{i-1}, b_{j-1})] = 0
\]

(9)

(15)

\[a_i = \Phi_1^{-1}(P_i)\]

(16)

\[b_j = \Phi_1^{-1}(P_j),\]

where \(P_{ij}\) is the observed proportion in cell \((i, j)\), and where \(P_i\) and \(P_j\) are observed cumulative marginal proportions of the table, i.e.,

(17)

\[P_i = \sum_{k=1}^{i} \sum_{j=1}^{r} P_{kj}\]

and

(18)

\[P_j = \sum_{i=1}^{s} \sum_{k=1}^{j} P_{ik}.\]
4. Variance/Covariance Estimates

Case 1

We denote the sample size by $N$, and introduce the notation $\mathbf{\theta}' = (\rho, a_1, a_2, \ldots, a_{s-1}, b_1, b_2, \ldots, b_{r-1})$. The matrix $I_{(\mathbf{\theta})}$ of expected second order derivatives of $l$ with respect to $\mathbf{\theta}$ is obtained from

$$[I_{(\mathbf{\theta})}]_{m,n} = N \sum_{i=1}^{s} \sum_{j=1}^{r} \frac{1}{\pi_{ij}} \left( \frac{\partial \pi_{ij}}{\partial \theta_m} \right) \left( \frac{\partial \pi_{ij}}{\partial \theta_n} \right)$$

[Tallis, 1962, p 348]. The derivatives within parenthesis in (19) are obtained from (8) and (10).

A large-sample estimate of the covariance matrix of $\mathbf{\theta}$ is therefore

$$V = I_{(\mathbf{\theta})}^{-1}$$
Case 2

We define \( P' = (P_{11}, P_{12}, \ldots, P_{1r}, P_{21}, \ldots, P_{2r}, \ldots, P_{s1}, \ldots, P_{sr}) \). The estimators are defined as the solution to the system

\[
F_1(\theta, P) = \frac{\partial l}{\partial \rho} \cdot \frac{1}{N} = 0
\]

(21)

\[
[F_2(\theta, P)]_i = a_i - \Phi_i^{-1}(P_i) = 0
\]

\[
[F_3(\theta, P)]_j = b_j - \Phi_i^{-1}(P_j) = 0
\]

Let \( F = (F_1 \quad F_2' \quad F_3') \).
The estimator $\theta$ is implicitly defined from $F(\theta, P) = 0$.

Since

$$\frac{\partial \theta}{\partial P} = - \left( \frac{\partial F}{\partial \theta} \right)^{-1} \frac{\partial F}{\partial P},$$

(22)

it follows from asymptotic theory that the large-sample covariance matrix for $\theta$ is,

$$\text{cov} (\theta) = \left( \frac{\partial F}{\partial \theta} \right)^{-1} \left( \frac{\partial F}{\partial P} \right) \Sigma \left( \frac{\partial F}{\partial P} \right)' \left( \frac{\partial F}{\partial \theta} \right)^{-1},$$

(23)

where $\Sigma$ is the covariance matrix of $P$, and where $\partial F/\partial P$ and $\partial F/\partial \theta$ are computed in the probability limit.
In (23), $\partial F/\partial \theta$ is structured as

$$
\begin{bmatrix}
\frac{\partial^2 l}{\partial \rho^2} & \frac{\partial^2 l}{\partial \rho \partial a} & \frac{\partial^2 l}{\partial \rho \partial b} \\
0 & I & 0 \\
0 & 0 & I
\end{bmatrix}
$$

(24)

The second order derivatives in (24) are given in the Appendix. Since $n_{ij} = N \cdot P_{ij}$, we get for the first row of $\partial F/\partial P$:

(25) \[ \frac{\partial F_1}{\partial P_{ij}} = \frac{1}{\pi_{ij}} [\phi_2(a_i, b_j) - \phi_2(a_{i-1}, b_j) - \phi_2(a_i, b_{j-1}) + \phi_2(a_{i-1}, b_{j-1})]. \]
Also, since $D[f^{-1}(y_0)] = 1/D(f(x_0))$, and since $\partial P_k/\partial P_{ij} = 1$, we get

(26) \[
\frac{\partial F_{2k}}{\partial P_{ij}} = \begin{cases} 
0 & \text{if } P_k \text{ does not contain } P_{ij}, \text{i.e. if } k < i \\
-rac{\partial}{\partial P_{ij}} [\Phi^{-1}(P_k)] &= -\frac{\partial \Phi_1^{-1}}{\partial P_k} \cdot \frac{\partial P_k}{\partial P_{ij}} \\
&= -\frac{1}{\frac{\partial}{\partial a_k} [\Phi_1(a_k)]} \\
&= \frac{1}{\phi_1(a_k)} \text{ otherwise.}
\end{cases}
\]
Similarly,

\[
\frac{\partial F_{3m}}{\partial P_{ij}} = \begin{cases} 
0 & \text{if } m < j \\
-\frac{1}{\phi_1(b_m)} & \text{otherwise.}
\end{cases}
\]

In \( \Sigma \), finally, the general element \( \sigma_{mn} \) is obtained from

\[
N\sigma_{mn} = \begin{cases} 
P_m (1 - P_m) & \text{if } m = n \\
-P_m P_n & \text{otherwise.}
\end{cases}
\]
5. Some Numerical Results

The formulae of the preceding section have been used in a computer program for maximum likelihood estimation of polychoric correlations, written by the author. The program reads a cross-table (or, at the user's option, raw data from which the cross-table is prepared). Starting values for the thresholds and the correlation are computed, and the program iteratively solves the likelihood equations by a Newton-Raphson algorithm. The program uses subroutines from the IMSL [1975] library, except for the bivariate normal distribution function, which was adopted from Kirk [1973]. The output consists of final threshold estimates, correlation estimate, variance/covariance matrix of the estimates, and the function value at the maximum. The latter may be used to test the fit of the model. Empty rows and/or columns of the table are deleted from the computations. The numerical results consist of the following steps:

(i) Some runs were made with the program where the cell frequencies are expected values, taken from some division of the bivariate normal surface. Thus, these runs are made with artificial population data. These runs may be used partly to check the numerical accuracy of the program, but first of all they give the expected variances and covariances of the estimates, in large samples. If the expected correlations between $\hat{\rho}$ and the thresholds are small this may serve as an indication that simultaneous estimation of $\rho$ and thresholds is unnecessary.

(ii) Some runs were made with Monte Carlo data, generated from the population data. These simulations were performed in order to check the theoretical large-sample results derived in Section 4.

(iii) Finally, some empirical data were analyzed.
Artificial Population Data

*Design.* The data were chosen according to a factorial design with the following parameters. The true correlation, \( \rho \), was chosen as .15, .50 and .85. The number of classes in the table, \( r = s \), was 2, 3, 5 and 7. The thresholds were placed so that, for each variable, the probabilities of the classes 1, \( \cdots \), \( s \) were the same as those of a binomial distribution with parameters \( P \) and \( n = s - 1 \). Parameter \( P \) was chosen so that the skewness of the distribution attained specified values \( \gamma \). This procedure was the same as that of Olsson [1979]. The combinations of skewnesses used were

\[
(0,0), \ (1, -1), \ (1, 0) \ \text{and} \ (1, 1).
\]
Results. The absolute difference between true correlation and the correlation computed by the program was in no case larger than $10^{-5}$, which implies reasonable accuracy of the program. For reasons of space, we do not exhibit all the variance/covariance matrices of the parameter estimates here. In Table 2 we only give one typical example.

For low values of $\rho$, the expected correlations between $\hat{\rho}$ and the thresholds are low; for $\rho = 0$ they are zero. For higher values of $\rho$, however, these expected correlations are in some cases more substantial, with a typical peak value of .20. This implies an increasing degree of dependence between correlation estimate and threshold estimates with increasing $\rho$. In the Monte Carlo data we shall study how much this means in practice.

In Table 3 we present the expected variance for $\hat{\rho}$ for the different combinations of parameters, given $N = 500$. For other values of $N$, the variances are proportional. The values are the same for both methods, to the given accuracy.
The case \( r = s = 2 \) was included in the computations with the following problem in mind: Given a set of skewed ordinal data, should the correlations be computed as (a) a polychoric correlation, or (b) a tetrachoric correlation from data dichotomized near the median?

It can be seen in Table 3, that the variances are uniformly smaller for the polychoric than for the tetrachoric correlation. Given highly skewed trichotomous data \((\gamma_1/\gamma_2 = 1./-1. \text{ or } 1./1.)\) the loss in efficiency if it is possible to dichotomize at the median is moderate, or may even be a small gain. In general, however, there seems to be a loss in efficiency of alternative (b) as compared to alternative (a) above.
TABLE 2

An Example of the Expected Covariance (in and below the diagonal) and Correlation (above the diagonal) Matrix for the Parameter Estimates. Artificial Population Data with \( r=s=5, \gamma_1=1, \gamma_2=0 \). and \( \rho=0.15 \)

<table>
<thead>
<tr>
<th></th>
<th>( \rho )</th>
<th>( a_1 )</th>
<th>( a_2 )</th>
<th>( a_3 )</th>
<th>( a_4 )</th>
<th>( b_1 )</th>
<th>( b_2 )</th>
<th>( b_3 )</th>
<th>( b_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \rho )</td>
<td>0.00281</td>
<td>0.1895</td>
<td>0.04680</td>
<td>0.05151</td>
<td>-0.00368</td>
<td>0.05022</td>
<td>0.02250</td>
<td>-0.02482</td>
<td>-0.04738</td>
</tr>
<tr>
<td>( a_1 )</td>
<td>0.00059</td>
<td>0.33913</td>
<td>0.20104</td>
<td>0.06310</td>
<td>0.02311</td>
<td>0.00822</td>
<td>0.00972</td>
<td>0.00761</td>
<td>0.00379</td>
</tr>
<tr>
<td>( a_2 )</td>
<td>0.00040</td>
<td>0.18730</td>
<td>0.02558</td>
<td>0.31104</td>
<td>0.11377</td>
<td>0.02717</td>
<td>0.03449</td>
<td>0.02914</td>
<td>0.01598</td>
</tr>
<tr>
<td>( a_3 )</td>
<td>0.00021</td>
<td>0.00277</td>
<td>0.00375</td>
<td>0.00569</td>
<td>0.36431</td>
<td>0.05186</td>
<td>0.07101</td>
<td>0.06473</td>
<td>0.03880</td>
</tr>
<tr>
<td>( a_4 )</td>
<td>-0.00001</td>
<td>0.00075</td>
<td>0.00102</td>
<td>0.00154</td>
<td>0.00315</td>
<td>0.06111</td>
<td>0.09513</td>
<td>0.09126</td>
<td>0.06003</td>
</tr>
<tr>
<td>( b_1 )</td>
<td>0.00023</td>
<td>0.00042</td>
<td>0.00038</td>
<td>0.00034</td>
<td>0.00030</td>
<td>0.00774</td>
<td>0.38355</td>
<td>0.17451</td>
<td>0.06619</td>
</tr>
<tr>
<td>( b_2 )</td>
<td>0.00007</td>
<td>0.00033</td>
<td>0.00032</td>
<td>0.00031</td>
<td>0.00030</td>
<td>0.00198</td>
<td>0.0343</td>
<td>0.45465</td>
<td>0.17455</td>
</tr>
<tr>
<td>( b_3 )</td>
<td>-0.00008</td>
<td>0.00026</td>
<td>0.00027</td>
<td>0.00029</td>
<td>0.00030</td>
<td>0.00090</td>
<td>0.00156</td>
<td>0.00343</td>
<td>0.38352</td>
</tr>
<tr>
<td>( b_4 )</td>
<td>-0.00022</td>
<td>0.00019</td>
<td>0.00022</td>
<td>0.00026</td>
<td>0.00030</td>
<td>0.00051</td>
<td>0.00090</td>
<td>0.00198</td>
<td>0.00774</td>
</tr>
</tbody>
</table>
TABLE 3

Expected Variances for \( r \) for the Artificial Population Data. The Values are the Same for Both Methods.

<table>
<thead>
<tr>
<th>( \rho )</th>
<th>( r )</th>
<th>0.0/0.</th>
<th>( \gamma_{1}/\gamma_{2} )</th>
<th>1.0/-1.</th>
<th>1.0/0.</th>
<th>1.0/1.</th>
</tr>
</thead>
<tbody>
<tr>
<td>.15</td>
<td>2</td>
<td>.00478</td>
<td>.00635</td>
<td>.00545</td>
<td>.00607</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>.00295</td>
<td>.00429</td>
<td>.00354</td>
<td>.00417</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>.00231</td>
<td>.00344</td>
<td>.00281</td>
<td>.00339</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>.00215</td>
<td>.00318</td>
<td>.00261</td>
<td>.00314</td>
<td></td>
</tr>
<tr>
<td>.50</td>
<td>2</td>
<td>.00329</td>
<td>.00494</td>
<td>.00382</td>
<td>.00405</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>.00194</td>
<td>.00313</td>
<td>.00238</td>
<td>.00272</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>.00144</td>
<td>.00238</td>
<td>.00182</td>
<td>.00218</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>.00131</td>
<td>.00215</td>
<td>.00166</td>
<td>.00201</td>
<td></td>
</tr>
<tr>
<td>.85</td>
<td>2</td>
<td>.00080</td>
<td>.00295</td>
<td>.00108</td>
<td>.00096</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>.00046</td>
<td>.00104</td>
<td>.00064</td>
<td>.00063</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>.00028</td>
<td>.00069</td>
<td>.00040</td>
<td>.00047</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>.00023</td>
<td>.00058</td>
<td>.00034</td>
<td>.00041</td>
<td></td>
</tr>
</tbody>
</table>
Monte Carlo Data

*Design.* The parameters for the Monte Carlo data were chosen using the same factorial design as above, except that \( r = 2 \) was excluded. For each combination of parameters, 10 samples of size 500 were generated, using the multinomial routine GGMUL of the IMSL [1975] library. Each sample was analyzed using both methods.

*Results.* The estimates of the correlation coefficients in the different samples are displayed in Tables 4 to 6, for \( s = 3, 5 \) and 7, respectively, along with the mean value and the variance in each cell. In most cases, the two methods produce estimates which are very similar. A closer inspection of Tables 4 to 6 reveals, that the differences between the methods does increase with increasing \( \rho \). For \( \rho = .15 \) the largest difference is \( 8 \times 10^{-4} \), for \( \rho = .50 \) it is \( 31 \times 10^{-4} \) and for \( \rho = .85 \) it is \( 43 \times 10^{-4} \). This is in agreement with the results given above. These values might give an indication of magnitude of the difference between the two methods.
The question of bias of the estimates may be studied in several ways. We have performed the following comparisons:

The mean value of the 10 sample values were tested against the corresponding true value. Both when the theoretical and when the sample variance were used, 4 of the 36 sample means were significantly different from the true value at the 5% level. Since all the significant values were higher than the corresponding true values, this might indicate a tendency towards a slight positive bias. The results were the same for both methods. The significant values have been indicated in Tables 4–6.

The variances among the 10 observations in each sample were tested against the corresponding theoretical variances, using $\chi^2$ with 9 d.f. Only one value was significant at the 5% level (see Tables 4–6).
### TABLE 4
Estimated Correlations in Samples Generated with the Indicated Values of True Correlations and Skewnesses
\( r=s=3 \), \( n=500 \)
The left column in each cell is the two-step estimate, the right column is the full maximum likelihood estimate.

<table>
<thead>
<tr>
<th>( \rho )</th>
<th>( \frac{\gamma_1}{\gamma_2} )</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>.15</td>
<td>0./0.</td>
<td>1./-1.</td>
<td>1./0.</td>
<td>1./1.</td>
</tr>
<tr>
<td></td>
<td>.1277</td>
<td>.1277</td>
<td>.0342</td>
<td>.0343</td>
</tr>
<tr>
<td></td>
<td>.2245</td>
<td>.2245</td>
<td>.1414</td>
<td>.1414</td>
</tr>
<tr>
<td></td>
<td>.2007</td>
<td>.2007</td>
<td>.1754</td>
<td>.1755</td>
</tr>
<tr>
<td></td>
<td>1.018</td>
<td>1.018</td>
<td>.1912</td>
<td>.1912</td>
</tr>
<tr>
<td></td>
<td>.0767</td>
<td>.0767</td>
<td>.2324</td>
<td>.2324</td>
</tr>
<tr>
<td></td>
<td>.2032</td>
<td>.2033</td>
<td>.1649</td>
<td>.1649</td>
</tr>
<tr>
<td></td>
<td>.0740</td>
<td>.0740</td>
<td>.1560</td>
<td>.1561</td>
</tr>
<tr>
<td></td>
<td>.1546</td>
<td>.1546</td>
<td>.1418</td>
<td>.1419</td>
</tr>
</tbody>
</table>

**\( \bar{x} \)**

\( .1533 \)  \( .1533 \)  \( .1553 \)  \( .1554 \)  \( .1132 \)  \( .1132 \)  \( .1412 \)  \( .1413 \)

**\( s^2 \)**

\( .0028 \)  \( .0028 \)  \( .0023 \)  \( .0023 \)  \( .0041 \)  \( .0041 \)  \( .0046 \)  \( .0046 \)

### TABLE 4

Estimated Correlations in Samples Generated with the Indicated Values of True Correlations and Skewnesses

$r=s=3$.  $n=500$

The left column in each cell is the two-step estimate, the right column is the full maximum likelihood estimate.

<table>
<thead>
<tr>
<th></th>
<th>.4840</th>
<th>.4846</th>
<th>.5327</th>
<th>.5325</th>
<th>.4583</th>
<th>.4586</th>
<th>.5285</th>
<th>.5257</th>
</tr>
</thead>
<tbody>
<tr>
<td>.50</td>
<td>.5344</td>
<td>.5345</td>
<td>.5107</td>
<td>.5105</td>
<td>.4714</td>
<td>.4713</td>
<td>.5513</td>
<td>.5487</td>
</tr>
<tr>
<td>.50</td>
<td>.5197</td>
<td>.5199</td>
<td>.4851</td>
<td>.4857</td>
<td>.5594</td>
<td>.5599</td>
<td>.4699</td>
<td>.4699</td>
</tr>
<tr>
<td>.50</td>
<td>.5326</td>
<td>.5329</td>
<td>.3915</td>
<td>.3915</td>
<td>.4450</td>
<td>.4447</td>
<td>.6112</td>
<td>.6102*</td>
</tr>
<tr>
<td>.50</td>
<td>.4892</td>
<td>.4896</td>
<td>.4162</td>
<td>.4161</td>
<td>.4914</td>
<td>.4915</td>
<td>.5604</td>
<td>.5622</td>
</tr>
<tr>
<td>.50</td>
<td>.5399</td>
<td>.5392</td>
<td>.4317</td>
<td>.4319</td>
<td>.5002</td>
<td>.5002</td>
<td>.4205</td>
<td>.4197</td>
</tr>
<tr>
<td>.50</td>
<td>.4681</td>
<td>.4678</td>
<td>.4023</td>
<td>.4025</td>
<td>.4748</td>
<td>.4748</td>
<td>.5474</td>
<td>.5484</td>
</tr>
<tr>
<td>.50</td>
<td>.5063</td>
<td>.5058</td>
<td>.5023</td>
<td>.5023</td>
<td>.4803</td>
<td>.4803</td>
<td>.4379</td>
<td>.4385</td>
</tr>
<tr>
<td>.50</td>
<td>.5156</td>
<td>.5152</td>
<td>.5412</td>
<td>.5420</td>
<td>.5877</td>
<td>.5876*</td>
<td>.4638</td>
<td>.4650</td>
</tr>
<tr>
<td>.50</td>
<td>.5427</td>
<td>.5424</td>
<td>.6187</td>
<td>.6188*</td>
<td>.5115</td>
<td>.5117</td>
<td>.5327</td>
<td>.5329</td>
</tr>
</tbody>
</table>

| x    | .5133 | .5132 | .4832 | .4833 | .4980 | .4981 | .5124 | .5121 |
| s²   | .0006 | .0006 | .0047 | .0047 | .0018 | .0018 | .0034 | .0033 |
TABLE 4
Estimated Correlations in Samples Generated with the
Indicated Values of True Correlations and Skewnesses
r=s=3.  n=500
The left column in each cell is the two-step estimate,
the right column is the full maximum likelihood estimate.

<table>
<thead>
<tr>
<th>.85</th>
<th>.8326</th>
<th>.8321</th>
<th>.8440</th>
<th>.8440</th>
<th>.8484</th>
<th>.8484</th>
<th>.8544</th>
<th>.8555</th>
</tr>
</thead>
<tbody>
<tr>
<td>.8446</td>
<td>.8449</td>
<td>.8781</td>
<td>.8781</td>
<td>.8400</td>
<td>.8402</td>
<td>.8968</td>
<td>.8958*</td>
<td></td>
</tr>
<tr>
<td>.8681</td>
<td>.8677</td>
<td>.8249</td>
<td>.8249</td>
<td>.8328</td>
<td>.8830</td>
<td>.8730</td>
<td>.8714</td>
<td></td>
</tr>
<tr>
<td>.8527</td>
<td>.8532</td>
<td>.8791</td>
<td>.8791</td>
<td>.8306</td>
<td>.8307</td>
<td>.8570</td>
<td>.8574</td>
<td></td>
</tr>
<tr>
<td>.8739</td>
<td>.8737</td>
<td>.8775</td>
<td>.8775</td>
<td>.8761</td>
<td>.8762</td>
<td>.8656</td>
<td>.8665</td>
<td></td>
</tr>
<tr>
<td>.8307</td>
<td>.8301</td>
<td>.8506</td>
<td>.8506</td>
<td>.8573</td>
<td>.8574</td>
<td>.8373</td>
<td>.8360</td>
<td></td>
</tr>
<tr>
<td>.8491</td>
<td>.8485</td>
<td>.8764</td>
<td>.8764</td>
<td>.8592</td>
<td>.8594</td>
<td>.8414</td>
<td>.8399</td>
<td></td>
</tr>
<tr>
<td>.8715</td>
<td>.8728</td>
<td>.8427</td>
<td>.8427</td>
<td>.8645</td>
<td>.8647</td>
<td>.8680</td>
<td>.8697</td>
<td></td>
</tr>
<tr>
<td>.8389</td>
<td>.8393</td>
<td>.8231</td>
<td>.8231</td>
<td>.8452</td>
<td>.8454</td>
<td>.8630</td>
<td>.8619</td>
<td></td>
</tr>
<tr>
<td>.8679</td>
<td>.8677</td>
<td>.8433</td>
<td>.8433</td>
<td>.8654</td>
<td>.8657</td>
<td>.8431</td>
<td>.8460</td>
<td></td>
</tr>
</tbody>
</table>

| x | .8530  | .8530  | .8540  | .8540  | .8570  | .8571  | .8600  | .8600  |
| s² | .0002 | .0002 | .0004 | .0004 | .0002 | .0002 | .0003 | .0003 |

Note: Underlined entries are significantly different from the true value (5% level) when the true variance is used. Starred (*) entries are significantly different from the true value (5% level) when the variance estimate is taken from the sample. The stars apply to both entries in each pair.
TABLE 5

Estimated Correlations in Samples Generated with the Indicated Values of True Correlations and Skewnesses. $r=s=5$. $n=500$

The left column in each cell is the two-step estimate, the right column is the full maximum likelihood estimate.

<table>
<thead>
<tr>
<th>$\gamma_1/\gamma_2$</th>
<th>0./0.</th>
<th>1./-1.</th>
<th>1./0.</th>
<th>1./1.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.2540 .2540*</td>
<td>.0656 .0657</td>
<td>.1762 .1763</td>
<td>.0325 .0326</td>
<td></td>
</tr>
<tr>
<td>.1449 .1449</td>
<td>.1865 .1866</td>
<td>.2676 .2678*</td>
<td>.1329 .1329</td>
<td></td>
</tr>
<tr>
<td>.1909 .1909</td>
<td>.1738 .1738</td>
<td>.1575 .1575</td>
<td>.0767 .0767</td>
<td></td>
</tr>
<tr>
<td>.1700 .1700</td>
<td>.1523 .1523</td>
<td>.1399 .1403</td>
<td>.1306 .1307</td>
<td></td>
</tr>
<tr>
<td>.2247 .2247</td>
<td>.0941 .0940</td>
<td>.2306 .2305</td>
<td>.1457 .1457</td>
<td></td>
</tr>
<tr>
<td>.1760 .1760</td>
<td>.1303 .1304</td>
<td>.2312 .2313</td>
<td>.1773 .1773</td>
<td></td>
</tr>
<tr>
<td>.1426 .1426</td>
<td>.0462 .0462</td>
<td>.1927 .1935</td>
<td>.1469 .1471</td>
<td></td>
</tr>
<tr>
<td>.1456 .1456</td>
<td>.2315 .2316</td>
<td>.1419 .1419</td>
<td>.2483 .2483</td>
<td></td>
</tr>
<tr>
<td>.1636 .1636</td>
<td>.1306 .1307</td>
<td>.1612 .1612</td>
<td>.2187 .2188</td>
<td></td>
</tr>
<tr>
<td>.2072 .2072</td>
<td>.1010 .1010</td>
<td>.1589 .1590</td>
<td>.1603 .1605</td>
<td></td>
</tr>
<tr>
<td>$\bar{x}$</td>
<td>.1820 .1820*</td>
<td>.1312 .1312</td>
<td>.1858 .1859*</td>
<td>.1470 .1471</td>
</tr>
<tr>
<td>$s^2$</td>
<td>.0012 .0012</td>
<td>.0029 .0029</td>
<td>.0017 .0017</td>
<td>.0035 .0035</td>
</tr>
</tbody>
</table>
TABLE 5

Estimated Correlations in Samples Generated with
the Indicated Values of True Correlations and Skewnesses.
r=s=5.  n=500
The left column in each cell is the two-step estimate,
the right column is the full maximum likelihood estimate.

<table>
<thead>
<tr>
<th></th>
<th>.50</th>
<th>.5574</th>
<th>.5587</th>
<th>.5080</th>
<th>.5085</th>
<th>.4949</th>
<th>.4956</th>
<th>.4008</th>
<th>.4020</th>
</tr>
</thead>
<tbody>
<tr>
<td>.50</td>
<td>.5161</td>
<td>.5159</td>
<td>.5026</td>
<td>.5029</td>
<td>.5040</td>
<td>.5044</td>
<td>.4922</td>
<td>.4927</td>
<td></td>
</tr>
<tr>
<td>.5187</td>
<td>.5174</td>
<td>.4987</td>
<td>.4988</td>
<td>.5708</td>
<td>.5696</td>
<td>.4799</td>
<td>.4804</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.5522</td>
<td>.5524</td>
<td>.4027</td>
<td>.4024</td>
<td>.4904</td>
<td>.4899</td>
<td>.5107</td>
<td>.5098</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.4844</td>
<td>.4841</td>
<td>.5244</td>
<td>.5251</td>
<td>.4569</td>
<td>.4563</td>
<td>.5023</td>
<td>.5028</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.4782</td>
<td>.4780</td>
<td>.4348</td>
<td>.4352</td>
<td>.5050</td>
<td>.5044</td>
<td>.4790</td>
<td>.4785</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.4690</td>
<td>.4689</td>
<td>.6096</td>
<td>.6095*</td>
<td>.5421</td>
<td>.5413</td>
<td>.5404</td>
<td>.5412</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.5420</td>
<td>.5433</td>
<td>.5222</td>
<td>.5223</td>
<td>.5292</td>
<td>.5292</td>
<td>.5102</td>
<td>.5113</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.5400</td>
<td>.5398</td>
<td>.5119</td>
<td>.5123</td>
<td>.5435</td>
<td>.5442</td>
<td>.5372</td>
<td>.5377</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$x$ | .5216 | .5217 | .5068 | .5071 | .5186 | .5184 | .4884 | .4888 |
$s^2$ | .0010 | .0011 | .0029 | .0029 | .0010 | .0010 | .0017 | .0017 |
### TABLE 5

Estimated Correlations in Samples Generated with the Indicated Values of True Correlations and Skewnesses. \( r=s=5 \). \( n=500 \)

The left column in each cell is the two-step estimate, the right column is the full maximum likelihood estimate.

<table>
<thead>
<tr>
<th></th>
<th>( r = .85 )</th>
<th>( r = .8851 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.8738 .8734</td>
<td>.8370 .8373</td>
</tr>
<tr>
<td>2</td>
<td>.8613 .8615</td>
<td>.8414 .8416</td>
</tr>
<tr>
<td>3</td>
<td>.8782 .8775</td>
<td>.8600 .8599</td>
</tr>
<tr>
<td>4</td>
<td>.8717 .8711</td>
<td>.8262 .8261</td>
</tr>
<tr>
<td>5</td>
<td>.8652 .8654</td>
<td>.8183 .8180</td>
</tr>
<tr>
<td>6</td>
<td>.8429 .8429</td>
<td>.8497 .8496</td>
</tr>
<tr>
<td>7</td>
<td>.8445 .8455</td>
<td>.8468 .8471</td>
</tr>
<tr>
<td>8</td>
<td>.8560 .8562</td>
<td>.8727 .8726</td>
</tr>
<tr>
<td>9</td>
<td>.8851 .8859</td>
<td>.8306 .8305</td>
</tr>
</tbody>
</table>

- \( \bar{x} = .8654 .8654 \) *
- \( s^2 = .0002 .0002 \) *

Note: Underlined entries are significantly different from the true value (5% level) when the true variance is used. Starred (*) entries are significantly different from the true value (5% level) when the variance estimate is taken from the sample. The stars apply to both entries in each pair.
**TABLE 6**

Estimated Correlations in Samples Generated with the Indicated Values of True Correlations and Skewnesses.  
\( r=s=7 \quad n=500 \)

The left column in each cell is the two-step estimate, the right column is the full maximum-likelihood estimate.

<table>
<thead>
<tr>
<th>( \rho )</th>
<th>0./0.</th>
<th>1./-1.</th>
<th>1./0.</th>
<th>1./1.</th>
</tr>
</thead>
<tbody>
<tr>
<td>.15</td>
<td>.1428</td>
<td>.1433</td>
<td>.1429</td>
<td>.1430</td>
</tr>
<tr>
<td></td>
<td>.1168</td>
<td>.1169</td>
<td>.2287</td>
<td>.2290</td>
</tr>
<tr>
<td></td>
<td>.1690</td>
<td>.1693</td>
<td>.1676</td>
<td>.1676</td>
</tr>
<tr>
<td></td>
<td>.1031</td>
<td>.1033</td>
<td>.2381</td>
<td>.2385</td>
</tr>
<tr>
<td></td>
<td>.1543</td>
<td>.1546</td>
<td>.2229</td>
<td>.2228</td>
</tr>
<tr>
<td></td>
<td>.2256</td>
<td>.2262</td>
<td>.1140</td>
<td>.1141</td>
</tr>
<tr>
<td></td>
<td>.1778</td>
<td>.1779</td>
<td>.1557</td>
<td>.1557</td>
</tr>
<tr>
<td></td>
<td>.1426</td>
<td>.1428</td>
<td>.2274</td>
<td>.2274</td>
</tr>
<tr>
<td></td>
<td>.1257</td>
<td>.1259</td>
<td>.1927</td>
<td>.1931</td>
</tr>
<tr>
<td></td>
<td>.2409</td>
<td>.2413*</td>
<td>.1105</td>
<td>.1105</td>
</tr>
</tbody>
</table>

\( \bar{x} \)

| .1599     | .1602     | .1801     | .1802     | .1476     | .1478     | .1584     | .1584     |

\( s^2 \)

| .0018     | .0018     | .0021     | .0021     | .0014     | .0014     | .0038     | .0038     |
TABLE 6

Estimated Correlations in Samples Generated with the Indicated Values of True Correlations and Skewnesses. r=s=7  n=500

The left column in each cell is the two-step estimate, the right column is the full maximum-likelihood estimate.

<table>
<thead>
<tr>
<th>.50</th>
<th>.4542</th>
<th>.4555</th>
<th>.5216</th>
<th>.5216</th>
<th>.5022</th>
<th>.5049</th>
<th>.5289</th>
<th>.5287</th>
</tr>
</thead>
<tbody>
<tr>
<td>.5037</td>
<td>.4829</td>
<td>.4829</td>
<td>.5083</td>
<td>.5097</td>
<td>.5072</td>
<td>.5066</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.4793</td>
<td>.3976</td>
<td>.3984*</td>
<td>.5126</td>
<td>.5148</td>
<td>.5429</td>
<td>.5399</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.5283</td>
<td>.5288</td>
<td>.4994</td>
<td>.4991</td>
<td>.5698</td>
<td>.5697</td>
<td>.4958</td>
<td>.4956</td>
<td></td>
</tr>
<tr>
<td>.5607</td>
<td>.5611</td>
<td>.4809</td>
<td>.4807</td>
<td>.4982</td>
<td>.4990</td>
<td>.6026</td>
<td>.6051*</td>
<td></td>
</tr>
<tr>
<td>.5314</td>
<td>.5335</td>
<td>.4220</td>
<td>.4222</td>
<td>.4688</td>
<td>.4692</td>
<td>.4949</td>
<td>.4947</td>
<td></td>
</tr>
<tr>
<td>.4905</td>
<td>.4908</td>
<td>.4707</td>
<td>.4712</td>
<td>.4978</td>
<td>.4983</td>
<td>.5658</td>
<td>.5653</td>
<td></td>
</tr>
<tr>
<td>.4223</td>
<td>.4226*</td>
<td>.4916</td>
<td>.4918</td>
<td>.4948</td>
<td>.4949</td>
<td>.5566</td>
<td>.5583</td>
<td></td>
</tr>
<tr>
<td>.4678</td>
<td>.4686</td>
<td>.4804</td>
<td>.4806</td>
<td>.4424</td>
<td>.4417</td>
<td>.5670</td>
<td>.5680</td>
<td></td>
</tr>
<tr>
<td>.5415</td>
<td>.5430</td>
<td>.4820</td>
<td>.4824</td>
<td>.5384</td>
<td>.5408</td>
<td>.4700</td>
<td>.4722</td>
<td></td>
</tr>
</tbody>
</table>

\[ \bar{x} = .4980 .4988 \quad .4729 .4731 \quad .5033 .5043 \quad \underline{5332 .5334} * \]

\[ s^2 = .0017 .0017 \quad .0012 .0012 \quad .0011 .0011 \quad .0015 .0015 \]
TABLE 6

Estimated Correlations in Samples Generated with the Indicated Values of True Correlations and Skewnesses. \( r=s=7, \ n=500 \)
The left column in each cell is the two-step estimate, the right column is the full maximum-likelihood estimate.

<table>
<thead>
<tr>
<th>.85</th>
<th>.8470</th>
<th>.8476</th>
<th>.8493</th>
<th>.8493</th>
<th>.8362</th>
<th>.8388</th>
<th>.8313</th>
<th>.8286</th>
</tr>
</thead>
<tbody>
<tr>
<td>.8718</td>
<td>.8728</td>
<td>.8418</td>
<td>.8421</td>
<td>.8523</td>
<td>.8528</td>
<td>.8438</td>
<td>.8434</td>
<td></td>
</tr>
<tr>
<td>.8323</td>
<td>.8334</td>
<td>.8916</td>
<td>.8915*</td>
<td>.8633</td>
<td>.8632</td>
<td>.8713</td>
<td>.8700</td>
<td></td>
</tr>
<tr>
<td>.8267</td>
<td>.8270</td>
<td>.8621</td>
<td>.8621</td>
<td>.8414</td>
<td>.8427</td>
<td>.8579</td>
<td>.8596</td>
<td></td>
</tr>
<tr>
<td>.8745</td>
<td>.8750</td>
<td>.8131</td>
<td>.8130</td>
<td>.8730</td>
<td>.8727</td>
<td>.8743</td>
<td>.8786</td>
<td></td>
</tr>
<tr>
<td>.8445</td>
<td>.8457</td>
<td>.8859</td>
<td>.8859</td>
<td>.8671</td>
<td>.8678</td>
<td>.8390</td>
<td>.8388</td>
<td></td>
</tr>
<tr>
<td>.8106</td>
<td>.8105*</td>
<td>.8545</td>
<td>.8549</td>
<td>.8237</td>
<td>.8279</td>
<td>.8548</td>
<td>.8555</td>
<td></td>
</tr>
<tr>
<td>( \underline{.8772} )</td>
<td>( \underline{.8769} )</td>
<td>( \underline{.9210} )</td>
<td>( \underline{.9210} )</td>
<td>( \underline{.8719} )</td>
<td>( \underline{.8712} )</td>
<td>( .8321 )</td>
<td>( .8326 )</td>
<td></td>
</tr>
<tr>
<td>.8215</td>
<td>.8228</td>
<td>.8502</td>
<td>.8503</td>
<td>.8713</td>
<td>.8711</td>
<td>.8568</td>
<td>.8582</td>
<td></td>
</tr>
<tr>
<td>.8506</td>
<td>.8520</td>
<td>.8585</td>
<td>.8585</td>
<td>.8285</td>
<td>.8289</td>
<td>.7953</td>
<td>.7967*</td>
<td></td>
</tr>
</tbody>
</table>

\[ \bar{x} \] | .8457 | .8464 | .8628 | .8629 | .8529 | .8537 | .8457 | .8462 |
| \( s^2 \) | \( .0005 \) | \( .0005 \) | \( .0008 \) | \( .0008 \) | \( .0003 \) | \( .0003 \) | \( .0005 \) | \( .0005 \) |

Note: Underlined entries are significantly different from the true value (5% level) when the true variance is used. Starred (*) entries are significantly different from the true value (5% level) when the variance estimate is taken from the sample. The stars apply to both entries in each pair.
The variances among the 10 observations in each sample were tested against the corresponding theoretical variances, using $\chi^2$ with 9 d.f. Only one value was significant at the 5% level (see Tables 4–6).

Each single correlation estimate was tested against the corresponding true value, both using the theoretical variance and the estimate from the sample. The significant values at the 5% level are indicated in Tables 4–6. Using the theoretical variance, 20 samples out of 360 (5.6%) were significant, of which 11 were over-, and 9 under-estimates. For the empirical variance estimates, 21 were significant (5.8%) for the full ML estimate, and 21 (5.8%) for the two-step estimate. It may be noted that the differences in variance estimates between the two methods are very small. An example of a generated crosstable, including population parameter values and estimates, is given in Table 7.
Analysis of Some Real Data

As a complement to the Monte Carlo results, we have also analyzed some real data with each of the two methods. The data consist of answers to nine five-step attitude items of very varying skewness (about from \(-3\) to \(+2\)). The data was kindly supplied by Dr. Thorleif Pettersson. The sample size was 329. In Table 8 we give the raw correlations, as well as correlations estimated with the two-step (TS) and full maximum likelihood (ML) methods.

The results do support the conclusions arrived at earlier, that the differences between the methods are quite small, but that they increase with increasing \(\rho\). The differences between the correlations estimated by our methods and the correlations computed as if the data had been on an interval scale (in the table denoted “Raw correlation”) are seen to be large.
A Generated Cross-table, and the Corresponding Parameter Values and Sample Estimates.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>13</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>69</td>
<td>113</td>
<td>22</td>
</tr>
<tr>
<td>3</td>
<td>41</td>
<td>132</td>
<td>104</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ρ</th>
<th>$a_1$</th>
<th>$a_2$</th>
<th>$b_1$</th>
<th>$b_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population value</td>
<td>.50</td>
<td>-1.70</td>
<td>-.31</td>
<td>-.67</td>
</tr>
<tr>
<td>Sample estimate</td>
<td>.49</td>
<td>-1.77</td>
<td>-.14</td>
<td>-.69</td>
</tr>
<tr>
<td>Estimated standard error</td>
<td>.048</td>
<td>.103</td>
<td>.056</td>
<td>.061</td>
</tr>
<tr>
<td>Variables</td>
<td>Raw correlation</td>
<td>ML-estimate</td>
<td>TS-estimate</td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>-----------------</td>
<td>-------------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>2 1</td>
<td>.470</td>
<td>.7277</td>
<td>.7215</td>
<td></td>
</tr>
<tr>
<td>2 1</td>
<td>.413</td>
<td>.5979</td>
<td>.5958</td>
<td></td>
</tr>
<tr>
<td>3 2</td>
<td>.390</td>
<td>.6267</td>
<td>.6240</td>
<td></td>
</tr>
<tr>
<td>4 1</td>
<td>-.084</td>
<td>-.2444</td>
<td>-.2412</td>
<td></td>
</tr>
<tr>
<td>4 2</td>
<td>-.001</td>
<td>-.0995</td>
<td>-.0990</td>
<td></td>
</tr>
<tr>
<td>4 3</td>
<td>-.459</td>
<td>-.6074</td>
<td>-.5899</td>
<td></td>
</tr>
<tr>
<td>5 1</td>
<td>.448</td>
<td>.7026</td>
<td>.6878</td>
<td></td>
</tr>
<tr>
<td>5 2</td>
<td>.364</td>
<td>.6202</td>
<td>.6140</td>
<td></td>
</tr>
<tr>
<td>5 3</td>
<td>.455</td>
<td>.6496</td>
<td>.6453</td>
<td></td>
</tr>
<tr>
<td>5 4</td>
<td>-.128</td>
<td>-.2752</td>
<td>-.2712</td>
<td></td>
</tr>
<tr>
<td>6 1</td>
<td>.414</td>
<td>.5898</td>
<td>.5880</td>
<td></td>
</tr>
<tr>
<td>6 2</td>
<td>.323</td>
<td>.5362</td>
<td>.5348</td>
<td></td>
</tr>
<tr>
<td>6 3</td>
<td>.540</td>
<td>.6583</td>
<td>.6523</td>
<td></td>
</tr>
<tr>
<td>6 4</td>
<td>-.313</td>
<td>-.4187</td>
<td>-.4116</td>
<td></td>
</tr>
<tr>
<td>6 5</td>
<td>.479</td>
<td>.7056</td>
<td>.6978</td>
<td></td>
</tr>
<tr>
<td>7 1</td>
<td>.355</td>
<td>.4912</td>
<td>.4908</td>
<td></td>
</tr>
<tr>
<td>7 2</td>
<td>.329</td>
<td>.5083</td>
<td>.5065</td>
<td></td>
</tr>
<tr>
<td>7 3</td>
<td>.431</td>
<td>.5353</td>
<td>.5319</td>
<td></td>
</tr>
<tr>
<td>7 4</td>
<td>-.172</td>
<td>-.2511</td>
<td>-.2478</td>
<td></td>
</tr>
<tr>
<td>7 5</td>
<td>.395</td>
<td>.6075</td>
<td>.6016</td>
<td></td>
</tr>
<tr>
<td>7 6</td>
<td>.556</td>
<td>.6531</td>
<td>.6470</td>
<td></td>
</tr>
<tr>
<td>8 1</td>
<td>.187</td>
<td>.2767</td>
<td>.2760</td>
<td></td>
</tr>
<tr>
<td>8 2</td>
<td>.081</td>
<td>.2021</td>
<td>.2013</td>
<td></td>
</tr>
<tr>
<td>8 3</td>
<td>.169</td>
<td>.2308</td>
<td>.2301</td>
<td></td>
</tr>
<tr>
<td>8 4</td>
<td>.015</td>
<td>.0170</td>
<td>.0170</td>
<td></td>
</tr>
<tr>
<td>8 5</td>
<td>.210</td>
<td>.3366</td>
<td>.3351</td>
<td></td>
</tr>
<tr>
<td>8 6</td>
<td>.202</td>
<td>.2533</td>
<td>.2525</td>
<td></td>
</tr>
<tr>
<td>8 7</td>
<td>.112</td>
<td>.1639</td>
<td>.1628</td>
<td></td>
</tr>
<tr>
<td>9 1</td>
<td>-.129</td>
<td>-.2519</td>
<td>-.2508</td>
<td></td>
</tr>
<tr>
<td>9 2</td>
<td>.049</td>
<td>-.0270</td>
<td>-.0269</td>
<td></td>
</tr>
<tr>
<td>9 3</td>
<td>-.340</td>
<td>-.4426</td>
<td>-.4385</td>
<td></td>
</tr>
<tr>
<td>9 4</td>
<td>.427</td>
<td>.5299</td>
<td>.5246</td>
<td></td>
</tr>
<tr>
<td>9 5</td>
<td>-.139</td>
<td>-.2751</td>
<td>-.2728</td>
<td></td>
</tr>
<tr>
<td>9 6</td>
<td>-.254</td>
<td>-.3312</td>
<td>-.3284</td>
<td></td>
</tr>
<tr>
<td>9 7</td>
<td>-.205</td>
<td>-.2775</td>
<td>-.2742</td>
<td></td>
</tr>
<tr>
<td>9 8</td>
<td>.087</td>
<td>.1055</td>
<td>.1050</td>
<td></td>
</tr>
</tbody>
</table>
6. Discussion

To summarize the results of the preceding section, it seems that the bias of the estimates is small, and that the variances are close to the theoretically derived ones for the data analyzed so far. Inspection of the cross-tables for the few extreme cases does suggest that bad estimates are more likely in tables where some marginals are small, i.e. where some expected cell frequencies are low. Since we have not studied this problem systematically, this statement must be taken in a loose sense, however.

For practical purposes, the differences between the two methods discussed here seem to be small, especially when the true correlation is small. This also applies for the variance estimates. It may be noted here, that the simpler variance estimates used by Martinson and Hamdan [1971] also turn out to be sufficiently accurate for practical work.
One problem which may have some practical importance is, that when several correlations are estimated, the full maximum likelihood estimate may lead to different threshold estimates for variable $x$ when $\rho_{xy}$ is estimated than when $\rho_{xz}$ is estimated. From a theoretical point of view, this is not entirely satisfactory. One solution is to use the two-step estimate, for which the thresholds are estimated from the marginals. A second solution would be to estimate the correlations for all variables simultaneously, including all the thresholds.

Another problem concerns the robustness of the methods. We have assumed here that underlying each response there is some latent variable which is normally distributed; an assumption, by the way, which is testable. In applications, such distributional assumptions are seldom exactly met. It might be a worthwhile task to examine to what degree departures from the assumption of normality has any effect on the correlation estimates.
Appendix: Some Details of the Derivations

We shall here derive the derivative of $F_1$ with respect to $\rho$, $a_i$, and $b_j$, computed at the probability limit.

\[(A1)\quad F_1 = \frac{1}{N} \sum \sum \frac{n_{ij}}{\pi_{ij}} [\phi(a_i, b_j) - \phi(a_{i-1}, b_j) - \phi(a_i, b_{j-1}) + \phi(a_{i-1}, b_{j-1})].\]

Since

\[(A2)\quad \frac{\delta}{\delta \rho} [\phi(u, v)] = \phi(u, v) \cdot \frac{uv(1 - \rho^2) + \rho(u^2 - 2\rho uv + v^2) - 2\rho(1 - \rho^2)}{(1 - \rho^2)^2} = g(u, v),\]

say, it follows that the derivative of $F_1$ with respect to $\rho$ may be written

\[(A3)\quad \frac{\delta F_1}{\delta \rho} = \frac{1}{N} \sum \sum \frac{n_{ij}}{\pi_{ij}} [g(a_i, b_j) - g(a_{i-1}, b_j) - g(a_i, b_{j-1}) + g(a_{i-1}, b_{j-1})]
- \frac{1}{N} \sum \sum \frac{n_{ij}}{\pi_{ij}^2} \cdot \frac{\delta \pi_{ij}}{\delta \rho} \cdot [\phi(a_i, b_j) - \phi(a_{i-1}, b_j) - \phi(a_i, b_{j-1}) + \phi(a_{i-1}, b_{j-1})].\]
Since, in the probability limit, \( n_{ij}/N \pi_{ij} = 1 \), this reduces to

\[
\text{plim}_{N \to \infty} \frac{\partial F_1}{\partial \rho} = \sum \sum [g(a_\sigma, b_j) - g(a_{i-1}, b_j) - g(a_\sigma, b_{j-1}) \\
+ g(a_{i-1}, b_{j-1})] - \sum \sum \frac{1}{P_{ij}} \left( \frac{\partial \pi_{ij}}{\partial \rho} \right)^2.
\]

In (A4), it is easily seen that the first sum is reduced to

\[
g(a_\sigma, b_\sigma) - g(a_\sigma, b_0) - g(a_\sigma, b_r) + g(a_\sigma, b_0) = 0.
\]

This is so because all other terms cancel, and because \( g \) is zero in all points containing \( \alpha_0 \), \( \alpha_\sigma \), \( b_0 \) and/or \( b_r \). Thus,

\[
\text{plim}_{N \to \infty} \frac{\partial F_1}{\partial \rho} = -\sum \sum \frac{1}{\pi_{ij}} \left( \frac{\partial \pi_{ij}}{\partial \rho} \right)^2.
\]
For the derivatives with regard to the thresholds, we may write

\[(A7) \quad \frac{\delta F_i}{\delta a_k} = \frac{1}{N} \sum \left\{ \frac{n_{ij}}{\pi_{ij}} \left[ \frac{\delta \phi(a_k, b_j)}{\delta a_k} - \frac{\delta \phi(a_k, b_{j-1})}{\delta a_k} \right] - \frac{n_{i+1,j}}{\pi_{i+1,j}} \left[ \frac{\delta \phi(a_k, b_j)}{\delta a_k} - \frac{\delta \phi(a_k, b_{j-1})}{\delta a_k} \right] \right\}

- \frac{1}{N} \sum \sum \frac{n_{ij}}{\pi_{ij}^2} \frac{\delta \pi_{ij}}{\delta a_k} \left[ \phi(a_i, b_j) - \phi(a_{i-1}, b_j) - \phi(a_i, b_{j-1}) + \phi(a_{i-1}, b_{j-1}) \right].\]

Since, as noted before,

\[\text{plim} \frac{1}{N} \frac{n_{ij}}{\pi_{ij}} = 1,\]

the two terms in the first sum cancel. Therefore,

\[(A8) \quad \text{plim} \frac{\partial F_i}{\partial a_k} = - \sum \sum \frac{1}{\pi_{ij}} \left( \frac{\partial \pi_{ij}}{\partial \rho} \right) \cdot \left( \frac{\partial \pi_{ij}}{\partial a_k} \right).\]

The result for \(\partial F_i/\partial b_m\) is analogous.
CALCULATION OF THE TETRACHORIC CORRELATION COEFFICIENT

D. R. Divgi

SYRACUSE UNIVERSITY

A new subroutine has been developed for calculating the tetrachoric correlation coefficient. Recent advances in computing inverse normal and bivariate normal distributions have been utilized. The iterative procedure is started with an approximation with an error less than ±.0135.

I am grateful to the Editor for valuable suggestions for improving the presentation.

Requests for reprints should be sent to D. R. Divgi, Test Scoring & Evaluation Services, Syracuse University, Syracuse, NY 13210.

0033-3123/79/0600-0169$00.75/0
© 1979 The Psychometric Society
The tetrachoric correlation coefficient is very useful for describing the relationship between two dichotomous variables, such as a pair of test items scored right or wrong. [Carroll, 1961; Lord & Novick, 1968] One may conceive of the observed dichotomous variables as manifestations of underlying psychological traits which are normally distributed. Then the tetrachoric correlation coefficient is the product-moment correlation between these traits.

Let the marginals $p_1$ and $p_2$ be the proportions of individuals with correct answers on Items 1 and 2 respectively. The corresponding standard scores in a normal distribution are defined by $F(h) = p_1$ and $F(k) = p_2$ where $F(z)$ is the area under the normal curve from $z$ to $\infty$:

$$F(z) = \frac{1}{\sqrt{2\pi}} \int_z^\infty \exp \left(- \frac{x^2}{2}\right) dx.$$  

(1)

If random variables $x$ and $y$ have a standard bivariate normal distribution with correlation $r$, the probability of $(x > h, y > k)$ is given by

$$L(h, k, r) = \frac{1}{2\pi \sqrt{1 - r^2}} \int_h^\infty \int_k^\infty \exp \left(- \frac{x^2 + y^2 - 2rxy}{2(1 - r^2)}\right) dx dy.$$  

(2)
When one wants to calculate the tetrachoric correlation between two items, the probability $L(h, k, r)$ is set equal to $p_{11}$, the proportion of individuals with correct answers on both items. Then one solves (2) for $r$ by using an approximate formula or by an interactive procedure involving numerical integration. Except in theoretical studies, the values of $p_1$, $p_2$, and $p_{11}$ are obtained from a finite sample, and hence the value of $r$ is subject to sampling fluctuations as well as to computational error.
The Algorithm

Accurate calculation of the tetrachoric correlation using a computer involves three basic steps. First one must calculate the standard scores $h$ and $k$ from the marginals. This is done by using one of the available approximations, followed by an iterative procedure if needed for greater accuracy.

Once $h$ and $k$ are available, one seeks a solution of

$$L(h, k, r) = p_{11}.$$ 

As this equation cannot be solved for $r$ analytically, the second step in the procedure is to compute an approximate value of the correlation. The better this initial value, the less computer time it takes to achieve the desired accuracy.

Finally one uses an iterative procedure which yields a sequence of values which are closer and closer to the correct one. Any such procedure involves, at each step, computation of $L$ using the current estimate of $r$. This calculation must be made accurately because error in computing $L$ leads to error in the estimate of the correlation. Moreover, other things being equal, faster calculation of the bivariate probability results in more rapid evaluation of the correlation. Therefore the method used to compute the probability function $L$ is an important part of one’s algorithm for the tetrachoric correlation. The tetrachoric series of Pearson (1900) can be used, but the number of terms required for reasonable accuracy becomes very large if $|r|$ is close to unity [Kirk, 1973]. Kirk used 8-point Gaussian integration and double precision arithmetic to evaluate the integral in (2). This procedure, too, loses accuracy as $r$ increases, with the error rising to $10^{-3}$ when $r = .99$. 
A faster algorithm has been prepared by making improvements at all three stages.

(a) The standard scores $h$ and $k$ are computed from the marginals using the approximation of Odeh and Evans [1974]. Let $q = \min(p_1, 1 - p_1)$ and $y = \{-2 \ln(q)\}^{1/2}$. Then the approximation is of the form

$$|h| \approx y + \frac{S_4(y)}{T_4(y)},$$

where $S_4$ and $T_4$ are fourth degree polynomials. The value of $h$ is negative if $p_1$ exceeds .5. The maximum error in the approximation is $1.8 \times 10^{-6}$ when single precision floating point arithmetic is used in computers with a 32-bit word.

(b) The initial estimate of the correlation is calculated with the approximation described below. The absolute error in this approximation has been found to be smaller than .0135 provided the smallest cell proportion is no less than .001. The accuracy of the approximation reduces the number of iterations required later. In addition, because iteration begins so close to the correct value, the iterative procedure always converges.
(c) The bivariate normal probability $L(h, k, r)$ is computed with a new algorithm by the author [Divgi, in press]. This algorithm, using single precision, uniformly provides an accuracy of $10^{-6}$ or better at any values of the arguments. Moreover it is faster than the 8-point quadrature used by Kirk [1973]. Each new estimate of the correlation is calculated from the previous one by the Newton-Raphson method:

$$
\hat{r}_{t+1} = \hat{r}_t - \frac{L(h, k, \hat{r}_t) - p_{11}}{L'(h, k, \hat{r}_t)},
$$

where $L'(h, k, r)$ is the derivative with respect to $r$. (For the formula for $L'$ and its justification see Pearson, 1900.) Iteration is stopped when the increment $(\hat{r}_{t+1} - \hat{r}_t)$ becomes smaller in absolute value than a specified criterion $\epsilon_r$, or if the difference $|L(h, k, r_t) - p_{11}|$ becomes smaller than a given limit $\epsilon_p$.

The algorithm has been coded in FORTRAN IV as a subroutine with six arguments. No printed output is produced by the subroutine. The input variables are $p_1$, $p_2$, and $p_{11}$; the computed values of $h$, $k$, and $\hat{r}$ are returned to the calling program. The algorithm was compared for speed with Kirk's [1973] subroutine using $\epsilon_r = 10^{-4}$ and $\epsilon_p = 0$. The new algorithm was found to be 2.5 times as fast as Kirk's, requiring .01 second per correlation on an IBM 370/155. Listing of the subroutine, including the function for bivariate normal probability, is available from the author.
The Initial Value

A familiar approximation to the tetrachoric correlation is Pearson’s [1900] cosine formula. Let \( p_{12} = p_1 - p_{11} \), \( p_{21} = p_2 - p_{11} \), and \( p_{22} = 1 - p_{11} - p_{21} - p_{12} \) be the proportions in the other cells of the \( 2 \times 2 \) table. Let

\[
R(h, k, r) = \frac{p_{11}p_{22}}{-p_{12}p_{21}}.
\]

The cosine approximation is

\[
r \approx \cos \left( \frac{\pi}{1 + \sqrt{R}} \right).
\]

This yields exact values when \( h = k = 0 \), but its error can exceed .5 if \( h, k \) are large. Several different approximations have been reviewed by Castellan [1966]. The most accurate of these is the one given by Camp [1934], but it is restricted to \( |r| < 0.8 \) and cannot be expressed in analytical form.
A better approximation can be found by starting with the form

\begin{equation}
(4a) \quad r = \cos \frac{\pi}{1 + R^\alpha}
\end{equation}

which is exact if

\begin{equation}
(4b) \quad \alpha(h, k, r) = \frac{\log \left( \frac{\pi - \arccos r}{\arccos r} \right)}{\log R}
\end{equation}

where \(\arccos\) is taken to lie in \([0, \pi]\). Once the proportions \(p_1, p_2,\) and \(p_{11}\) are known, \(R\) can be calculated. If some function of the proportions provides an approximate value of \(\alpha\), this can be substituted in \((4a)\) to obtain an approximation for \(r\). An approximation containing 10 unknown numbers has been obtained for \(\alpha\) (and hence for \(r\)). The initial values were found by least squares and were further adjusted so as to minimize the maximum error in the correlation.
the correlation. The final approximation was as follows:

\[ r \approx \cos \left\{ \frac{\pi}{(1 + R^\alpha)} \right\}, \]

where

\[ \alpha = A + B \left\{ -1 + \frac{1}{1 + C \log R - D^2} \right\} \]

with

\[ A = \frac{.5}{1 + (h^2 + k^2) \left\{ .12454 - .27102 \left[ 1 - \frac{h}{(h^2 + k^2)^{1/2}} \right] \right\}}, \]

\[ B = \frac{.5}{1 + (h^2 + k^2) \left\{ .82281 - 1.03514 \frac{k}{(h^2 + k^2)^{1/2}} \right\}}, \]

\[ C = .07557 h + (h - k)^2 \left( \frac{.51141}{h + 2.05793} - \frac{.07557}{h} \right), \]

\[ D = k \left( .79289 + \frac{4.28981}{1 + 3.30231 h} \right). \]
These expressions are valid when $h \geq k \geq 0$ with $\alpha = .5$ when $h = k = 0$. Because of the symmetry properties of the functions (which are easy to derive from those of the cell proportions), other situations are handled as follows. Replace $h$ in the above equations with $\max(|h|, |k|)$, $k$ with $\min(|h|, |k|)$, and multiply $D$ by the signs of $h$ and $k$. The error of the approximation is less than .0135 provided each cell proportion exceeds .001. This condition should be satisfied in nearly all experimental data, and if it is not, effects of sampling errors are likely to be much larger. The accuracy of the approximation can be improved further if the lower limit on cell proportions is raised to, say, .01.

This approximation has been used to provide the starting value in the computer program. It may be sufficiently accurate as it stands for some uses. All the operations involved are available on hand calculators.
REFERENCES


*Manuscript received 1/26/78*

*First revision received 7/3/78*

*Final version received 1/16/79*
Identification of Binary Response Models

CHARLES F. MANSKI*

© 1988 American Statistical Association
Journal of the American Statistical Association
September 1988, Vol. 83, No. 403, Theory and Methods
1. INTRODUCTION

1.1 The Threshold-Crossing Model

The following "threshold-crossing" model of binary response has been applied extensively in economics, medicine, and other fields. An observable binary outcome \( z \) is assumed to be determined by an observable random variable \( x \) with domain \( X \subseteq \mathbb{R}^K \) and by an unobservable random variable \( u \) with domain \( \mathbb{R}^1 \). The pair \((x, u)\) act on \( z \) through the sign of a linear function \( x\beta + u \), where \( \beta \) is a \( K \)-dimensional parameter. That is,

\[
z = 1[x\beta + u \geq 0], \tag{1}
\]

where \( 1[*] \) is the function that takes the value 1 if the logical condition inside the brackets is satisfied and 0 otherwise.
In economic applications, $z$ usually indicates a utility-maximizing decision maker's observable choice between two alternatives. Then $x\beta + u$ is the difference between the utilities of these alternatives. In medical research, $z$ typically denotes an observable binary indicator of health status. Here, $x\beta + u$ is a latent continuous variable determining health status.

Assume that it is possible to observe realizations of $(z, x)$ drawn at random from the population. Accepting the threshold-crossing model as correct, we may distinguish between "reduced-form" and "structural" analyses of these binary response data. The objective of reduced-form analysis is inference on the set of conditional response probabilities $P_{1|x} \equiv \{\Pr(z = 1 \mid x), x \in X\}$. Structural analysis seeks to learn about the random threshold-crossing process determining these probabilities.
For each $x$ in $X$, the probability of the event $[z = 1]$ is

$$P_{1|x} = \int 1[x\beta + u \geq 0] \, dF_{u|x},$$

(2)

where $F_{u|x}$ denotes the distribution function for $u$ conditional on the realization of $x$. So the objective of structural analysis is inference on $\beta$ and on the conditional distributions $F_{u|X} \equiv (F_{u|x}, \, x \in X)$. 
1.2 Structural Analysis

This article studies the central necessary condition for meaningful structural analysis, namely identification. Before going any further, we need to ask why one might be interested in a structural interpretation of binary response data. Reduced-form analysis is simpler. In principle, the response probabilities $P_{1|x}$ may be estimated by nonparametric regression of $z$ on $x$ without reference to a threshold-crossing process. Or convenient parametric models can be fit to $(z, x)$ data, as is done in log-linear analysis. Why then should one attempt inference on a hypothetical random response function $x\beta + u$?
Here, as in other contexts where the distinction between reduced form and structure arises, we can point to two motives for structural analysis. First, one may have a scientific interest in learning about the process yielding binary outcomes, whether a decision process determining choice or a physiological process determining health status. If so, one necessarily seeks a structural interpretation of the data.
Second, one may want to make predictions of the binary outcome $z$ conditional on realizations of the observable variables $x$. The best predictor of $z$ given $x$ is always some function of the response probability $P_{1|x}$. For example, the best predictor under squared loss is $P_{1|x}$. Under absolute loss, the best predictor is $1[P_{1|x} \geq .5]$. Nevertheless, structural analysis has important applications in prediction problems. These include the following:

1. *Estimation of Best Predictors.* Information restricting $(\beta, F_{u|x})$ may imply restrictions on best predictors. If so, estimation of best predictors subject to these restrictions may be more precise and/or more tractable than unconstrained estimation.

2. *Extrapolation.* Reduced-form analysis can reveal the best predictor of $z$ given $x$ only in regions of $X$ that have positive probability mass. Structural analysis can make possible prediction at every point in the domain of the structural model.
1.3 Identification

Most research on binary response has considered specific estimators and tests. It is common to propose a method for inference on some structural feature and then to seek conditions under which the method has appealing statistical properties. Consideration of identification is often only a way station on the road to proof of consistency.
Although method-oriented research is productive, it risks obscuring the basic facts that unite or separate different methods. A newly proposed procedure may be touted as a competitor to others when it actually is non-comparable, in the sense of presuming distinct prior restrictions on the structural process. Or one method may be claimed superior to another when it really is nested, in the sense of requiring stronger prior information for it to be useful. The systematic study of identification can expose the foundations of binary response analysis by clarifying what information is needed to make different methods work.
Let $F_x$ denote the probability distribution of $x$. Analysis of identification assumes that the response probabilities are known almost everywhere (a.e.) $F_x$ and that some information about the pair $(\beta, F_{u|x})$ is available. In this idealized setting, we seek to determine the set of threshold-crossing processes, consistent with the available information, that satisfy Equation (1) for almost every realization of $x$. For now, we adopt the following definition.
Definition. Let $\Phi$ denote the space of all probability distributions on the real line. Let $\Phi^X$ be the Cartesian product of $\Phi$ over the space $X$. Let it be known that $(\beta, F_{u|x}) \in \Omega,$ where $\Omega$ is a given subset of $R^K \times \Phi^X$. For each $(b, G_{u|x})$ in $R^K \times \Phi^X$, let

$$X(b, G_{u|x}) = \{x \in X : P_{1|x} \neq \int 1[xb + u \geq 0] \ dG_{u|x}\}.$$

(3)

We say that $(\beta, F_{u|x})$ is identified relative to $(b, G_{u|x})$ if either $(b, G_{u|x}) \not\in \Omega$ or $\Pr[x \in X(b, G_{u|x})] > 0$. 
Let $\Omega_0 \subset \Omega$ denote the set of pairs $(b, G_{u|X})$ relative to which $(\beta, F_{u|X})$ is not identified. Let $c(\ast, \ast)$ be a function from $R^K \times \Phi^X$ onto a space $C$. Consider any $d \in C$. We say that $c(\beta, F_{u|X})$ is identified relative to $d$ if there exists no $(b, G_{u|X}) \in \Omega_0$ such that $c(b, G_{u|X}) = d$. We say that $c(\beta, F_{u|X})$ is identified if $c(b, G_{u|X}) = c(\beta, F_{u|X})$ for all $(b, G_{u|X}) \in \Omega_0$.

In Section 4, this definition will be extended in certain respects.
Note that the definition of identification refers to the lower probability $\Pr[x \in X(b, G_u|x)]$ rather than to a probability. This technical nicety covers the case in which the set $X(b, G_u|x)$ is not measurable. Henceforth we shall ignore the question of measurability and speak only of probabilities. Where the set under consideration is, in fact, not measurable, it should always be possible to rephrase the discussion in terms of lower probabilities.
1.4 Identification and Consistent Estimation

The most that reduced-form analysis of binary response data can achieve is to reveal the response probabilities arbitrarily closely, in certain senses, as the sample size increases. So identification analysis begins where reduced-form analysis ends. As is well known, identification of a function $c(\beta, F_{u|x})$ is necessary but not sufficient for consistent estimability of that function. Identification requires that $c(\beta, F_{u|x})$ be distinguishable from $d \in C$, $d \neq c(b, F_{u|x})$, given knowledge of $P_{1|x}$, a.e. $F_x$. Consistent estimation requires that neighborhoods of $c(\beta, F_{u|x})$ be identified from knowledge of neighborhoods of the response probabilities. Thus the gap between identification and consistent estimability is "smoothness." Verification that an estimation problem is smooth in an appropriate sense is often a tedious, unenlightening task, which this article avoids. Various consistency results in the literature, however, will be noted.
1.5 The Determinants of Identification

Our definition of identification emphasizes that identification is not an all-or-nothing proposition. The available information may not be sharp enough to identify \((\beta, F_{u|x})\) relative to all other elements of \(R^K \times \Phi^X\) but may suffice to identify interesting features of the structural process. In general, the identified functions of \((\beta, F_{u|x})\) depend on two factors, namely (a) the available information restricting \((\beta, F_{u|x})\) and (b) the subsets of \(X\) having positive probability mass under \(F_x\). These determinants of identification cannot be analyzed in isolation from one another. Rather, it is their interaction that matters. An example shows this well.
Example. The following alternative pairs of specifications for factors (a) and (b) identify $\beta$ up to scale:

1. For each $x$ in $X$, $F_{u|x}$ is the same known, continuous, strictly increasing distribution function with median 0.
2. There exists no proper linear subspace of $R^K$ having probability 1 under $F_x$.

or

1. For each $x$ in $X$, the median of $F_{u|x}$ equals 0.
2. There exists no proper linear subspace of $R^K$ having probability 1 under $F_x$. Moreover, at least one component of $x$ has everywhere positive Lebesgue density conditional on the other components.

The first pair of conditions is assumed in the logit and probit models. The second pair was investigated in Manski (1975, 1985). Moving from the first to the second pair, one weakens the prior structure placed on $F_{u|x}$ at the cost of strengthening the condition satisfied by $F_x$. 
1.6 Information Restricting \((\beta, F_{u|x})\)

One would like to know the consequences for identification of all conceivable combinations of factors (a) and (b). But since there does not appear to be any effective way to conduct a complete identification analysis, this article examines a set of leading cases.

The following assumptions are maintained throughout.

*Maintained Assumptions.* The model \(z = 1[x\beta + u \geq 0]\) correctly describes the determination of \(z\) by \((x, u)\). The first component of \(x\) is constant; that is, \(x_1 = 1\). No information restricting the value of \(\beta\) is available. For each \(x\) in \(X\), the distribution function \(F_{u|x}\) is continuous and strictly increasing on \(R^1\).
These assumptions are standard in the literature. Moreover, the model $z = 1[x\beta + u \geq 0]$ is observationally equivalent to a seemingly more general class of models of the form $z = 1[f(x, u) \geq 0]$. See Appendix A.

Given the maintained assumptions, we ask questions of the following kind: If the distributions $F_{u|x}$ are known to have specified properties, what conditions on $F_x$ are necessary and/or sufficient to identify given functions of $(\beta, F_{u|x})$? In particular, we study identification under the following five assumptions.
Assumption MI (mean independence). \( \mathbb{E}(u \mid x) = 0 \ \forall \ x \in X. \)

Assumption QI (quantile independence). For a given \( \alpha \in (0, 1), \Pr(u < 0 \mid x) = \alpha \ \forall \ x \in X. \)

Assumption IS (index sufficiency). (a) \( F_{u \mid x} = F_{u \mid x\beta} \ \forall \ x \in X. \) (b) For each \( x \in X, \) the distribution \( F_{u \mid x\beta} \) is absolutely continuous with respect to Lebesgue measure. For each \( v \in R^1, \) let \( \phi(\ast \mid v) \) be the density function of \( u, \) conditional on the event \([x\beta = v]. \) For each \( u \in R^1, \phi(u \mid \ast) \) is differentiable. For each \( \eta \in R^1, \)

\[
\frac{\partial \int_{\eta}^{\infty} \phi(u \mid v) \, du}{\partial v} = \int_{\eta}^{\infty} \frac{\partial \phi(u \mid v)}{\partial v} \, du.
\]

Assumption SI (statistical independence). There exists a distribution \( F_0 \) on \( R^1 \) such that \( F_{u \mid x} = F_0 \ \forall \ x \in X. \) \( \Pr(u < 0) = .5. \)

Assumption KD (known distribution). There exists a known distribution \( F_0 \) on \( R^1 \) such that \( F_{u \mid x} = F_0 \ \forall \ x \in X. \) \( \Pr(u < 0) = .5. \)
Each of these distributional assumptions is prominent in the literature. Moreover, collectively they demonstrate well the dependence of identification on prior knowledge.

Note that, in general, the pair \((\beta, F_{u|X})\) can be identified at most up to location and scale normalizations. For our analysis, it is convenient not to make an explicit scale normalization. It is most convenient to vary the location normalization with the distributional assumption. Thus Assumption MI sets the mean of \(u\) to 0; QI makes the \(\alpha\) quantile of \(u\) 0; SI and KD set the median of \(u\) to 0; IS makes no location normalization. The interpretation of the intercept parameter \(\beta_1\) varies with the location normalization.
1.7 Organization of the Article

Considering the foregoing restrictions on $F_{u|x}$ in turn, each of Sections 2–6 presents a proposition giving conditions necessary and/or sufficient for identification of $\beta$ relative to any $b \neq \beta$. Some of the propositions also treat the more general question of identification of $(\beta, F_{u|x})$ relative to a specified $(b, G_{u|x})$. Following the propositions are corollaries giving conditions on $F_x$ that identify interesting functions of $(\beta, F_{u|x})$.

The propositions are proved in the text. Proofs of the corollaries are in Appendix B. For ease of reference, the conditions on $F_x$ that will be used are collected here.
**Condition X1.** There exists no proper linear subspace of $R^K$ having probability 1 under $F_x$.

**Condition X2.** There exists a $\mu > 0$ such that $\Pr(|x_k| < \mu) = 1$ ($k = 1, \ldots, K$). There exists a $\lambda > 0$ such that $|x\beta| \geq \lambda$, a.e. $F_x$.

**Condition X3.** There exists at least one $k \in [2, \ldots, K]$ such that $\beta_k \neq 0$ and such that, for almost every value of $w \equiv (x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_K)$, $\Pr[x_k \in (a_1, a_2) | w] > 0$ for all open intervals $(a_1, a_2) \subset R^1$. Without loss of generality, let $k = K$.

**Condition X4.** There exists an open set $Y_0 \subset R^{K-1}$ such that $\Pr[(x_2, \ldots, x_K) \in Y_1] > 0$ for all open sets $Y_1 \subset Y_0$. 
Condition X5.  \( F_x \) is multinomial with positive probability on each of the \( M \) elements of \( X^M \equiv (x^1, \ldots, x^M) \subset X. \ x^m \beta \neq 0 \ (m = 1, \ldots, M) \).

Condition X6.  There exists at least one \( k \in [2, \ldots, K] \) such that \( \beta_k \neq 0 \) and such that, for almost every value of \( w \equiv (x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_K) \), the following holds: There exists \( r_w \in R^1 \) such that either \( \Pr[x_k \in (a_1, a_2) \mid w] > 0 \) for all open intervals \( (a_1, a_2) \subset (-\infty, r_w) \) or \( \Pr[x_k \in (a_1, a_2) \mid w] > 0 \) for all open intervals \( (a_1, a_2) \subset (r_w, \infty) \). Without loss of generality, let \( k = K \).
2. MEAN INDEPENDENCE

Consider the familiar problem of inference on a linear model when realizations of the pair \((x\beta + u, x)\) are observed. It is well known that if \(u\) is mean independent of \(x\), then \(\beta\) is identified, provided only that \(F_x\) does not have all of its mass in any proper linear subspace of \(R^K\).

It is natural to ask whether the knowledge that \(u\) is mean independent of \(x\) retains any identifying power in the binary response setting where only realizations of \((1[x\beta + u \geq 0], x)\) are observed. In fact, it does not. This simple negative result was noted in Manski (1985) and is shown in Proposition 1.
Proposition 1. Let Assumption MI hold. Let $b \in R^K$. Then $\beta$ is not identified relative to $b$.

Proof. For each $x \in X$, the maintained assumption that $F_{u|x}$ is continuous and strictly increasing implies that $0 < P_{1|x} < 1$. There exists a continuous, strictly increasing distribution $G_{u|x}$ such that $\int udG_{u|x} = 0$ and $\int 1[u \geq -xb] dG_{u|x} = P_{1|x}$. So $(\beta, F_{u|x})$ is not identified relative to $(b, G_{u|x})$. 
3. QUANTILE INDEPENDENCE

In linear model analysis, the assertion that $u$ is mean independent of $x$ is often motivated only by a vague belief that the "central tendency" of $u$ does not vary with $x$. Use of the mean to express central tendency rather than, say, the median is rationalized by a belief that it does not much matter how one defines central tendency formally.
In binary response analysis, the sense in which one asserts $u$ to be independent of $x$ matters enormously. Whereas a mean independence restriction does not allow structural analysis, a quantile independence restriction does. In this section, we examine identification given the knowledge that a specified quantile of $F_{u|x}$ is independent of $x$. No further restrictions are imposed on the distribution of $u$. If, in fact, other quantiles are also independent of $x$, this information is ignored.
Proposition 2 gives the basic identification result, which elaborates on ideas developed in Manski (1985).

**Proposition 2.** Let Assumption QI hold. Let $b \in \mathbb{R}^K$. Define

$$Q_b \equiv [x \in X : xb < 0 \leq x\beta \cup x\beta < 0 \leq xb].$$

Then $\beta$ is identified relative to $b$ iff $\Pr(x \in Q_b) > 0$. 
Proof. We first show sufficiency. By the threshold-crossing model,

\[ P_{1|X} \geq 1 - \alpha \iff \Pr(u \geq -x\beta \mid x) \geq 1 - \alpha. \quad (4) \]

Under Assumption QI,

\[ \eta \leq 0 \iff \Pr(u \geq \eta \mid x) \geq 1 - \alpha. \quad (5) \]

It follows from (4) and (5) that

\[ x\beta \geq 0 \iff P_{1|X} \geq 1 - \alpha. \quad (6) \]
For each $x \in Q_\beta$, $(xb < 0 \cap P_{1|x} \geq 1 - \alpha)$ or $(xb \geq 0 \cap P_{1|x} < 1 - \alpha)$. Hence there exists no $G_{u|x}$ such that $\int 1[u \geq 0] \, dG_{u|x} = 1 - \alpha$ and $\int 1[u \geq -xb] \, dG_{u|x} = P_{1|x}$. Therefore, $\beta$ is identified relative to $b$ if $\Pr(x \in Q_\beta) > 0$.

To prove that $\Pr(x \in Q_\beta) > 0$ is necessary for identification, consider $x$ in $X - Q_\beta$. Define $G_{u|x}$ to be any strictly increasing distribution such that $\int 1[u \geq 0] \, dG_{u|x} = 1 - \alpha$ and $\int 1[u \geq -xb] \, dG_{u|x} = P_{1|x}$. Then $(b, G_{u|x})$ yields the same response probability as does $(\beta, F_{u|x})$ and $G_{u|x}$ satisfies Assumption QI. It follows that $\beta$ is not identified relative to $b$ if $\Pr(x \in X - Q_\beta) = 1$. 
3.1 Identification of $\beta$ up to Scale

Given that Assumption QI does not set the scale of $(\beta, F_{u|x})$, it should be the case that $\beta$ is not identified relative to its scalar multiples. Proposition 2 verifies this.

Now let us fix scale by selecting a norm $\|\|$ on $R^K$. By Proposition 2, Assumption QI identifies $\beta/\|\beta\|$ iff $F_x$ places positive probability on enough subsets of $X$ to make $\Pr(x \in Q_b) > 0$ for all $b$ such that $b/\|b\| \neq \beta/\|\beta\|$. (We adopt the following convention: $b = 0 \Rightarrow b/\|b\| = 0$.) The following results extend findings of Manski (1985).
Corollary 1 (Prop. 2). Let Assumption QI and Condition X2 hold. Then $\beta$ is not identified relative to $b$ if $|b_k - \beta_k| < \lambda/K\mu \ [k \in (1, \ldots, K)]$.

Corollary 2 (Prop. 2). Let Assumption QI and Conditions X1 and X3 hold. Then $\beta$ is identified relative to all $b$ such that $b/\|b\| \neq \beta/\|\beta\|$.
Corollary 1 shows that if quantile independence is the only restriction on $F_{u|x}$ and if $\beta \neq 0$, then $F_x$ having full rank does not suffice to identify $\beta/\|\beta\|$. In general, $\beta/\|\beta\|$ is not identified if $F_x$ is multinomial or if the response probabilities are always either above or below $1 - \alpha$.

Corollary 2 shows that if $K \geq 2$, then $\beta/\|\beta\|$ is identified if $F_x$ has full rank and places positive probability on a rich collection of subsets of $X$, in the sense of Condition X3. The corollary leaves open the possibility that $\beta/\|\beta\|$ is identified under conditions weaker than X1 and X3. In fact, X1 is essential (see Sec. 6). It appears that X3 can be weakened only in minor ways.
3.2 Identification of $\text{sgn}(\beta)$

Research to date has focused on the problem of estimating $\beta/\|\beta\|$. But useful structural analysis remains possible when $\beta$ is not identified up to scale. It is often of interest to learn the sign of some component of $\beta$, say the $K$th component. Corollary 3 gives a condition on $F_x$ sufficient to identify $\text{sgn}(\beta_K)$. 
Corollary 3 (Prop. 2). Let Assumption QI hold. For $x \in X$, let $w \equiv (x_1, \ldots, x_{K-1})$. Let $W$ be the domain of $w$. For $b \in R^K$, define

$$W_b \equiv [w \in W : \Pr(xb < 0 \mid w) > 0 \cap \Pr(xb > 0 \mid w) > 0].$$

(a) Suppose that $\beta_K \neq 0$. Then $\text{sgn}(\beta_K)$ is identified if $\Pr(w \in W_\beta) > 0$. (b) Suppose that $\beta_K = 0$. Then $\beta_K$ is identified if $\Pr(w \in W_b) > 0$ for all $b \in R^K$ such that $b_K \neq 0$.

To interpret this corollary, consider the case in which $x = (1, x_2)$. Thus $w = 1$ and $x_2 = \beta_1 + \beta_2 x_2$. Part (a) states that if $\beta_2 \neq 0$, then $\text{sgn}(\beta_2)$ is identified if $\Pr(x_2 < -\beta_1/\beta_2) > 0$ and $\Pr(x_2 > -\beta_1/\beta_2) > 0$. Part (b) requires more of $F_x$. In particular, $\beta_2 = 0$ is identified if $\Pr(x_2 < \eta) > 0$ and $\Pr(x_2 > \eta) > 0$ for every $\eta$ in $(-\infty, \infty)$. 
3.3 Applications to Prediction

In Section 1, we noted that structural analysis may be of use in predicting binary outcomes. Let $p \in [0, 1]$ denote a predictor of $z$. Suppose that one seeks the value of $p$ that minimizes expected loss under the $\alpha$-absolute loss function

$$L_\alpha(p, z) \equiv \alpha |p - z| \cdot 1[p < z]$$

$$+ (1 - \alpha) |p - z| \cdot 1[p > z].$$

(7)

Knowing that $u$ is $\alpha$-quantile independent of $x$ facilitates solution of this prediction problem. The following corollary gives the result.
Corollary 4 (Prop. 2). Let $\alpha \in (0, 1)$. Let $x \in X$. Then

$$1[P_1|_x \geq 1 - \alpha] = \arg \min_{p \in [0, 1]} E[L_\alpha(p, z) \mid x]. \quad (8)$$

Let Assumption QI hold. Then

$$1[x_\beta \geq 0] = \arg \min_{p \in [0, 1]} E[L_\alpha(p, z) \mid x]. \quad (9)$$
The result in (8) is a special case of the general fact that the best predictor of a random variable under $\alpha$-absolute loss is its $\alpha$ quantile. Conditional on $x$, the $\alpha$ quantile of the Bernoulli $z$ is $1[P_{1|x} \geq 1 - \alpha]$. The result in (9) says that if $u$ is $\alpha$-quantile independent of $x$, then $1[x\beta \geq 0]$ gives an alternative expression for the best predictor of $z$. This follows directly from (6) and (8).
Use of the structural predictor $1[x\beta \geq 0]$ rather than the reduced-form predictor $1[P_{1|x} \geq 1 - \alpha]$ has at least two advantages. First, structural prediction requires only that one estimate $\beta$ well enough to determine $\text{sgn}(x\beta)$; the response probability $P_{1|x}$ need not be estimated. Second, the structural predictor $1[x\beta \geq 0]$ can be used to predict $z$ conditional on any realization of $x$. In contrast, the reduced-form predictor $[P_{1|x} \geq 1 - \alpha]$ is computable only for $x$ in the support of $F_x$. 
3.4 Median Independence and Conditional Symmetry

Quantile independence restrictions sometimes make researchers uncomfortable. The assertion that a given quantile of $u$ does not vary with $x$ may lead one to ask: Why this quantile but not others? In the absence of a persuasive answer, one may feel compelled to adopt an extreme position: Either no quantiles of $u$ are independent of $x$ or all are. On reflection, though, one may feel that the former position understates the available information and the latter overstates it.
In some applications, a satisfactory representation of the available information is given by the assertion that the distributions $F_{u|x}$ are all symmetric about a common center. Let the center of symmetry be set equal to 0 to normalize location. Then the restriction is

$$
\int 1[u \geq -\eta] \, dF_{u|x} = 1 - \int 1[u \geq \eta] \, dF_{u|x},
$$

$$\forall \, \eta \in R^1, \forall \, x \in X. \quad (10)$$

Conditional symmetry implies median independence. But it does not restrict the manner in which other quantiles of $u$ vary with $x$. 
It is natural to ask whether the conditional symmetry restriction has identifying power beyond that of median independence. In binary response analysis the answer is negative. Informally, the reason is because binary response data impose on the threshold model the restrictions \( P_{1|x} = \int 1[u \geq -x\beta] \, dF_{u|x} \) but are uninformative regarding \( \int 1[u \geq x\beta] \, dF_{u|x} \). Formally, it suffices to reconsider Proposition 2 with the assumption of conditional symmetry replacing median independence. The necessity part of the proof continues to hold if the distributions \( G_{u|x} (x \in X) \) are chosen to be symmetric around 0 rather than simply to have median 0.
3.5 Estimation

Considering the case of median independence, Manski (1975) introduced the maximum score method for estimation of $\beta/\|\beta\|$. That article gave a partial proof of consistency, provided that $F_x$ has everywhere positive Lebesgue density. Manski (1985) proved strong consistency of maximum score estimation, provided only that the assumptions of Corollary 2 hold.
4. INDEX SUFFICIENCY

In reduced-form analyses of binary response, it is common to assume that the response probability varies with $x$ only through the filter of a scalar “index” $x\beta$. That is, there exists a $\beta \in R^K$ and function $h(*)$ mapping $R^1$ into $[0, 1]$ such that

$$P_{1|x} = h(x\beta) \quad \forall \ x \in X. \quad (11)$$

The form of $h(*)$ is usually specified. Several authors, however, have proposed methods for estimating $\beta$ when $h(*)$ is known only to satisfy regularity conditions.
The threshold-crossing model generates response probabilities of the form of (11) if part (a) of Assumption IS is satisfied, that is, if the index $x\beta$ is a sufficient statistic for the distribution $F_{u|x}$. Then

$$h(x\beta) = \int 1[u \geq -x\beta] \, dF_{u|x\beta}. \quad (12)$$

Index sufficiency does not restrict the distribution $F_{u|x\beta}$, so the form of $h(*)$ is not constrained.
The restrictions on response probabilities implied by index sufficiency differ fundamentally from those implied by quantile independence. As we have seen, quantile independence restricts the level of the response probability at each \( x \in X \); that is, \( x \beta \geq 0 \Leftrightarrow P_{1|x} \geq 1 - \alpha \). In contrast, index sufficiency restricts the relative sizes of the response probabilities at certain pairs of points \( (x, \xi) \in X \times X \). Part (a) of Assumption IS is equivalent to the statement

\[
x \beta = \xi \beta \Rightarrow P_{1|x} = P_{1|\xi}.
\]

Note that the converse of (13) need not hold.
4.1 Restrictions on Pairs of Response Probabilities

To study identification under distributional assumptions that restrict pairs of response probabilities, we need a technically trivial but conceptually important lemma.

**Identification Lemma.** $P_{1|x}$ is known a.e. $F_x$ iff $(P_{1|x}, P_{1|\xi})$ is known a.e. $F_x \times F_x$.

**Proof.** Assume that $P_{1|x}$ is known on $X^0 \subset X$ and not on $X - X^0$. Then $(P_{1|x}, P_{1|\xi})$ is known on $X^0 \times X^0$ and not on $(X \times X) - (X^0 \times X^0)$. Under the product measure $F_x \times F_x$, $Pr[(x, \xi) \in X^0 \times X^0] = Pr(x \in X^0)^2$. Hence $Pr(x \in X^0) = 1 \iff Pr[(x, \xi) \in X^0 \times X^0] = 1$.

Now assume that $(P_{1|x}, P_{1|\xi})$ is known on $X^{00} \subset X \times X$ and not on $(X \times X) - X^{00}$. Knowledge of a pair implies knowledge of each element in the pair. So the set $X^{00}$ must have the form $X^0 \times X^0$ for some $X^0 \subset X$. Hence $Pr[(x, \xi) \in X^{00}] = 1 \iff Pr(x \in X^0) = 1$. 
With this lemma we can extend the definition of identification given in Section 1 to cover situations in which distributional assumptions restrict pairs of response probabilities. In Section 1, we said that \((\beta, F_{u|x})\) is identified relative to \((b, G_{u|x})\) if either \((b, G_{u|x}) \notin \Omega\) or if \(\Pr[x \in X(b, G_{u|x})] > 0\). Henceforth, the following extended definition holds.
**Extended Definition.** For each \((b, G_{u|X})\) in \(R^K \times \Phi^X\), let

\[
XX(b, G_{u|X}) \equiv \{(x, \xi) \in X \times X : (P_{1|x}, P_{1|\xi}) \\
\neq \left(\int 1[xb + u \geq 0] \ dG_{u|x}, \int 1[\xi b + u \geq 0] \ dG_{1|\xi}\right)\}.
\]

(14)

We say that \((\beta, F_{u|X})\) is identified relative to \((b, G_{u|X})\) if either \((b, G_{u|X}) \notin \Omega\) or if \(\Pr[(x, \xi) \in XX(b, G_{u|X})] > 0\).
4.2 Local Linearity

We can now examine identification. Part (a) of Proposition 3 imposes only Assumption IS(a). The result is very limited identifying power. Part (b) adds the smoothness condition of Assumption IS(b). This implies that $P_{1|x}$ behaves locally as a linear function of $\beta$.

Proposition 3. (a) Let Assumption IS(a) hold. Let $b \in R^K$. Define

$$I_b \equiv [(x, \zeta) \in X \times X: xb = \zeta b \cap P_{1|x} \neq P_{1|\zeta}].$$

Then $\beta$ is identified relative to $b$ iff $\Pr[(x, \zeta) \in I_b] > 0$.

(b) Let Assumption IS hold. Then $\partial P_{1|x}/\partial x' = [dh(x\beta)/dv]\beta$. 
Proof. (a) The proof follows directly from (13) and the Identification Lemma. (b) Assumption IS implies that

\[ P_{1|x} = h(x\beta) \equiv \int 1[u \geq -x\beta] \, dF_{u|x\beta} \]

\[ = \int_{-x\beta}^{\infty} \phi(u \mid x\beta) \, du. \]  \hspace{1cm} (15)

It follows that \( h(*) \) is differentiable with

\[ dh(x\beta)/dv \equiv \partial \left[ \int_{-x\beta}^{\infty} \phi(u \mid x\beta) \, du \right] / \partial v \]

\[ = \phi(u \mid -x\beta) + \int_{-x\beta}^{\infty} [\partial \phi(u \mid x\beta) / \partial v] \, du. \]  \hspace{1cm} (16)
By the chain rule,

\[ \partial P_{1|x}/\partial x' = \left[ dh(x\beta)/dv \right] \left[ \partial (x\beta)/\partial x' \right] \]

\[ = \left[ dh(x\beta)/dv \right] \beta. \] (17)

Part (a) of Proposition 3 is essentially a negative result. \( I_b \) is a subset of \( [(x, \xi) \in X \times X: xb = \xi b \cap x \neq \xi] \). So a necessary condition for \( \beta \) to be identified relative to \( b \) is that the event \( [xb = \xi b \cap x \neq \xi] \) have positive probability under \( F_x \times F_x \). Suppose that \( (x_2, \ldots, x_K) \) has Lebesgue density on \( R^{k-1} \). Then the event \( [xb = \xi b \cap x \neq \xi] \) has positive probability only if \( (b_2, \ldots, b_K) = 0 \). So one can hope to identify \( \beta \) only relative to \( b \) of the form \( (b_1, 0, \ldots, 0) \).
Assumption IS(a) is weak because it operates only on those \((x, \xi)\) pairs that satisfy \(x\beta = \xi\beta\). It seems plausible that more positive identification results should emerge if (13) can be extended to a statement that \(P_{1|x}\) must be close to \(P_{1|\xi}\) whenever \(x\beta\) is close to \(\xi\beta\). Part (b) of Assumption IS makes this possible. The local linearity shown in part (b) of Proposition 3 permits identification of the slope parameters \((\beta_2, \ldots, \beta_K)\) up to scale and sign. Corollary 1 gives the result.
**Corollary 1 (Prop. 3).** Let Assumption IS and Condition X4 hold. Assume that there exist a $\xi \in X$ and a $k \in [2, \ldots, K]$ such that $(\xi_2, \ldots, \xi_K) \in Y_0$, $\partial P_{1|\xi}/\partial x_k \neq 0$, and $\partial P_{1|\xi}/\partial x'$ is continuous at $\xi$. Then $(\beta_2, \ldots, \beta_K)/\|\beta_2, \ldots, \beta_K\|$ is identified up to sign.

This positive finding assumes that $x$ has a rich support, in the sense of Condition X4. It may be that this condition can be weakened and identification retained. But identification must require that the distribution of $x$ not be multinomial. If $F_x$ is multinomial, knowledge of $P_{1|x}$ (a.e. $F_x$) reveals nothing about the derivatives $\partial P_{1|x}/\partial x'$. So the local linearity found in Proposition 3 has no identifying power.
4.3 Applications to Prediction

Knowledge that Assumption IS holds can facilitate prediction in at least two ways. First, it reduces the task of estimating the response probabilities \( P_{1|*} \), which map \( X \) into \([0, 1]\), to the smaller problem of estimating \( \beta \) and the function \( h(*) \), which maps \( R^1 \) into \([0, 1]\). Second, it offers at least modest opportunities for extrapolation. Suppose that \( \xi \in X \) is not in the support of \( F_x \) but \( \xi \beta \) is "close" to \( x \beta \), where \( x \) is such that \( P_{1|x} \) and \( \partial P_{1|x}/\partial x' \) are estimable. \( \xi \) need not be close to \( x \). Part (b) of Proposition 3 implies that

\[
P_{1|\xi} \approx P_{1|x} + (\xi - x) \partial P_{1|x}/\partial x'.
\]

(18)

So the information that Assumption IS holds can, in principle, make it possible to predict \( z \) conditional on \( \xi \). Note, however, that Assumption IS does not provide a criterion for evaluation of the accuracy of the approximation (18).
4.4 Nontrivial Sufficiency

The most prominent class of models satisfying index sufficiency is that in which \( u \) is statistically independent of \( x \). But identification under statistical independence is easily analyzed directly, without reference to index sufficiency (see Sec. 5). Thus Proposition 3 and its corollary are of applied interest primarily for inference on models in which the index \( x\beta \) is nontrivially sufficient for \( F_{u|x} \).
**Example.** In the conditional Poisson model for count data, a random variable $z^*$ is distributed Poisson with mean $x\beta$. This model can be written in the linear form $z^* = x\beta + u$, where $F_{u|x}$ is the Poisson$(x\beta)$ distribution with realizations translated to the left by $x\beta$ units. Suppose that $z^*$ is not observable but one can observe whether at least $z_0$ counts occur, where $z_0$ is a given positive integer. Thus one observes the binary response $z = 1[z^* \geq z_0] = 1[x\beta + u - z_0 \geq 0]$. The result is a threshold model in which $x\beta$ is minimally sufficient for $F_{u|x}$. 
4.5 Estimation

Ruud (1986), Stoker (1986), Ichimura (1987), and Powell, Stock, and Stoker (1986) considered estimation of \((\beta_2, \ldots, \beta_K)\) up to scale and sign. These papers do not work explicitly with the threshold model of binary response. Rather, they assume directly that (11) holds and that \(h(*)\) is a smooth function.
5. STATISTICAL INDEPENDENCE

Statistical independence is a rather strong restriction on the distributions $F_{u|x}$. On the one hand, $u$ is statistically independent of $x$ iff $u$ is $\alpha$-quantile independent for every $\alpha \in (0, 1)$. On the other hand, statistical independence is equivalent to stating that $x\beta$ is sufficient for $x$ and that $h(*)$ of (11) is continuous and strictly increasing. Statistical independence implies Assumption IS(b) if the distribution of $u$ is absolutely continuous.

The fact that statistical independence subsumes quantile independence and index sufficiency opens two routes to the study of identification. We shall follow the quantile independence route. The results reported subsequently resemble those given in Section 3.
5.1 Identification of \((\beta, F_0)\)

Corollary 2 (Prop. 2) reported that if one quantile of \(u\) is known to be independent of \(x\), Conditions X1 and X3 identify \(\beta\) up to scale. If all of the quantiles of \(u\) are independent of \(x\), these conditions on \(F_x\) identify \((\beta, F_0)\) up to scale. This result follows directly from Corollary 2 (Prop. 2); hence we label it Corollary 5.

**Corollary 5 (Prop. 2).** Let Assumption SI and Conditions X1 and X3 hold. Then \((\beta, F_0)\) is identified up to scale.
5.2 Identification of $\beta$

The foregoing result does not address the possibility that interesting functions of $(\beta, F_0)$ may be identified under conditions weaker than X1 and X3. Proposition 4 provides the necessary and sufficient condition for identification of $\beta$ relative to any $b$.

Proposition 4. Let Assumption SI hold. Let $b \in R^K$. Define

$$Q_b \equiv [x \in X: xb < 0 \leq x\beta \cup x\beta < 0 \leq xb]$$

$$R_b \equiv [(x, \xi) \in X \times X: (x - \xi)b < 0 \leq (x - \xi)\beta$$

$$\cup (x - \xi)\beta < 0 \leq (x - \xi)b].$$

Then $\beta$ is identified relative to $b$ iff

$$\Pr(x \in Q_b) + \Pr[(x, \xi) \in R_b] > 0.$$
Proof. Proposition 2 proved that \( \Pr(x \in Q_b) > 0 \) suffices to identify \( \beta \) relative to \( b \). Under Assumption SI, \( \Pr[(x, \xi) \in R_b] > 0 \) also suffices. To see this, observe that under SI,

\[
x \beta \geq \xi \beta \iff P_{1|x} \geq P_{1|\xi} \quad \forall \ (x, \xi) \in X \times X. \quad (19)
\]

Let \( (x, \xi) \in R_b \). Then either \((xb < \xi b \cap P_{1|x} \geq P_{1|\xi})\) or \((xb \geq \xi b \cap P_{1|x} < P_{1|\xi})\). It follows that \((P_{1|x}, P_{1|\xi}) \neq (\int 1[u \geq -xb] \, dG, \int 1[u \geq -\xi b] \, dG)\) for all probability distributions \( G \). Hence \( \beta \) is identified relative to \( b \) if \( \Pr[(x, \xi) \in R_b] > 0 \).

Now let us show that \( \beta \) is not identified relative to \( b \) if \( \Pr(x \in Q_b) = \Pr[(x, \xi) \in R_b] = 0 \). Define \( S_b \equiv [(x, \xi) \in X \times X: x \notin Q_b \cap \xi \notin Q_b \cap (x, \xi) \notin R_b] \). By this definition,

\[
(x \beta \geq 0 \iff xb \geq 0) \\
\cap (\xi \beta \geq 0 \iff \xi b \geq 0) \cap (x \beta \geq \xi \beta \iff xb \geq \xi b)
\]

\[
\forall (x, \xi) \in S_b. \quad (20)
\]
It is possible to find a continuous, strictly increasing distribution $G$ such that $\int 1[u \geq 0] \, dG = .5$ and $(P_{1|x}, P_{1|z}) = (\int 1[u \geq -xb] \, dG, \int 1[u \geq -\xi b] \, dG) \forall (x, \xi) \in S_b$. It follows that $(\beta, F_0)$ is not identified relative to $(b, G)$ if $\Pr[(x, \xi) \in S_b] = 1$. But $\Pr[(x, \xi) \in S_b] = 1$ iff $\Pr(x \in Q_b) = \Pr[(x, \xi) \in R_b] = 0$.

Comparison of Propositions 2 and 4 makes explicit how the identifying power of a single quantile independence restriction compares with that of statistical independence. Under Assumption QI, $\beta$ is identified relative to $b$ iff $\Pr(x \in Q_b) > 0$. Under SI, $\beta$ is identified relative to $b$ iff $\Pr(x \in Q_b) > 0$ or $\Pr[(x, \xi) \in R_b] > 0$. 
Three corollaries applying to Proposition 4 will be presented. These corollaries parallel closely Corollaries 1, 2, and 3 of Proposition 2, but the identification results reported here are stronger.

**Corollary 1 (Prop. 4).** Let Assumption SI and Condition X5 hold. Define

\[
\lambda_1 \equiv \min[|x^m \beta|; m = 1, \ldots, M], \\
\lambda_2 \equiv \min[|(x^m - x^n)\beta|; m, n = 1, \ldots, M; m \neq n], \\
\mu_1 \equiv \max[|x_k^m|; m = 1, \ldots, M; k = 1, \ldots, K], \\
\mu_2 \equiv \max[|x_k^m - x_k^n|; m, n = 1, \ldots, M; k = 1, \ldots, K].
\]

Then \( \beta \) is not identified relative to any \( b \) such that

\[
|b_k - \beta_k| < \min[\lambda_1/K\mu_1, \lambda_2/K\mu_2], \quad k = 1, \ldots, K.
\]
Corollary 2 (Prop. 4). Let Assumption SI and Conditions X1 and X6 hold. Then \((\beta_2, \ldots, \beta_K)/\|\beta_2, \ldots, \beta_K\|\) is identified.

Corollary 3 (Prop. 4). Let Assumption SI hold. For \((x, \xi) \in X \times X\), let \(w \equiv (x_1, \ldots, x_{K-1})\) and \(d \equiv (x_1 - \xi_1, \ldots, x_{K-1} - \xi_{K-1})\). Let \(W\) and \(D\) be the domains of \(w\) and \(d\). For \(b \in \mathbb{R}^K\), define

\[
W_b \equiv \{w \in W : \Pr(xb < 0 \mid w) > 0 \cap \Pr(xb > 0 \mid w) > 0\},
\]

\[
D_b \equiv \{d \in D : \Pr((x - \xi)b < 0 \mid d) > 0 \cap \Pr((x - \xi)b > 0 \mid d) > 0\}.
\]

(a) Suppose that \(\beta_K \neq 0\). Then \(\text{sgn}(\beta_K)\) is identified if \(\Pr(w \in W_\beta) + \Pr(d \in D_\beta) > 0\). (b) Suppose that \(\beta_K = 0\). Then \(\beta_K\) is identified if \(\Pr(w \in W_b) + \Pr(d \in D_b) > 0\) for all \(b \in \mathbb{R}^K\) such that \(b_K \neq 0\).
Corollary 1 shows that if $F_x$ is multinomial, then $\beta$ is not identified with respect to all $b$ in a neighborhood of $\beta$. The present result is not as negative as was Corollary 1 for Proposition 2. There, $\beta$ was found to be unidentified locally whenever $F_x$ satisfies Condition X2. Here, we assume Condition X5, which implies X2.
Corollary 2 shows that the slope components of $\beta$ are identified up to scale if the distribution of $x_K$ conditional on $w = (x_1, \ldots, x_{K-1})$ places positive probability on every open subset of a half-line of $R^1$. The endpoint and direction of this half-line can vary with $w$. This condition on $F_x$ is weaker than that assumed in Corollary 2 for Proposition 2. The earlier result, however, proved identification up to scale of all of the components of $\beta$, including the intercept.

Corollary 3 gives a condition on $F_x$ that suffices to identify the sign of a coefficient. The condition imposed here is weaker than that assumed in Corollary 3 for Proposition 2.
5.3 Applications to Prediction

At a minimum, knowing that $u$ is statistically independent of $x$ makes possible all of the applications of structural prediction previously shown possible given quantile independence and index sufficiency. At best, statistical independence identifies $(\beta, F_0)$ up to scale. If so, evaluation of the expression $\int 1[u \geq -x\beta] \, dF_0$ enables best prediction of binary response at any $x \in X$, under any loss function.
5.4 Estimation

The methods for estimation of $\beta$ discussed in the preceding sections can all be applied if statistical independence holds. In addition, Cosslett (1983) considered maximum likelihood estimation of ($\beta$, $F_0$). He assumed essentially the conditions of Corollary 5 (Prop. 2) and proved consistency. Other estimators were studied by Han (1987) and Klein and Spady (1987).
6. KNOWN DISTRIBUTION

Most empirical research has assumed that $u$ is statistically independent of $x$ with a known, continuously increasing distribution $F_0$. Relative to Assumption SI, Assumption KD reduces the task of inferring $(\beta, F_0)$ to one of inferring $\beta$ alone. The consequence is that full rank of $F_x$ is necessary and sufficient for identification. Proposition 5 gives the well-known result.

**Proposition 5.** Let Assumption KD hold. (a) Let $b \in \mathbb{R}^K$. Then $\beta$ is identified relative to $b$ iff $\Pr(x\beta \neq xb) > 0$. (b) $(\beta, F_0)$ is identified iff Condition X1 holds.
Proof. (a) Let \( x \in X \). Given that \( F_0 \) is continuous and strictly increasing, \( \int 1[u \geq -x\beta] \, dF_0 = \int 1[u \geq -xb] \, dF_0 \Leftrightarrow \beta \leq x\beta = xb \). The result follows. (b) \( \Pr[x((\beta - b) \neq 0] > 0 \ \forall \ b \neq \beta \) iff no proper linear subspace of \( \mathbb{R}^K \) has probability 1 under \( F_x \).
7. RELATED IDENTIFICATION PROBLEMS

This article has developed several themes. We have distinguished between reduced-form and structural analysis of binary response data and have indicated the applications of structural analysis to prediction. We have emphasized that mean independence, quantile independence, and index sufficiency are separate distributional restrictions and have shown the implications of each for identification. We have seen that successively more positive identification results emerge as the information available on the distribution of $u$ is strengthened from quantile independence to statistical independence to statistical independence with a known distribution.
We have organized the analysis by asking: If the distributions $F_{u|x}$ are known to have specified properties, what conditions on $F_x$ are necessary and/or sufficient to identify given functions of $(\beta, F_{u|x})$? A researcher given a sample design may wish to ask the converse question: If $F_x$ has specified properties, what assumptions on $F_{u|x}$ identify structural parameters? The corollaries to Propositions 1–5 provide some answers. The corollaries report findings for designs in which $x$ is discrete, in which $x$ has at least one continuous component, and in which $x$ is absolutely continuous.
Two themes not developed here have been taken up elsewhere, at least in part. Hsieh, Manski, and McFadden (1985) and Manski (1986) examined identification of binary response models from response-based samples. Under response-based sampling, reduced-form analysis reveals only the "odds ratios" \( [P_{1|x}/P_{0|x}]/[P_{1|\xi}/P_{0|\xi}] \), a.e. \( F_x \times F_x \). The response probabilities themselves are not revealed. Nevertheless, positive identification results have been obtained under Assumptions QI and SI, as well as KD.
Anderson (1970) and Manski (1987) investigated identification from binary response panel data, which make possible identification in the presence of unrestricted, unobserved person-specific effects. Cross-sectional data have no identifying power in this setting.

Throughout this article, we have been careful not to attach any significance to the response function $x\beta + u$ beyond its sign. Assume now that $x\beta + u$ is, in principle, observable. Then observation of binary response data is interpretable as the consequence of censoring. In fact, binary censoring is the most extreme censoring process that preserves any information about the latent continuous response.
Having studied identification under binary censoring, one would like to compare the findings reported here with those that obtain under less extreme forms of censoring. In particular, it would be of interest to complement the present study of identification with one performed under the assumption that \( \max[0, x\beta + u] \) is observed. Much research has been devoted to the development of estimators and tests for application to such censored data. But the underlying identification question has not been studied systematically.
APPENDIX A: LINEAR LATENT RESPONSE MODELS

Sign-Preserving Transformations

The statement \( z = 1[x\beta + u \geq 0] \) really means that \( z = 1[f(x, u) \geq 0] \) for some function \( f(\ast, \ast) \) that can be made linear by a "sign-preserving" transformation.
**Definition.** A mapping \( \theta : R^1 \rightarrow R^1 \) is said to be sign-preserving if, for every \( \eta \in R^1, \eta \geq 0 \iff \theta(\eta) \geq 0 \).

In the threshold-crossing model of binary response, all sign-preserving transformations of the latent variable are equivalent, in that they imply precisely the same binary outcomes. Hence a latent response function \( f(*, *) \) may be considered linear if it has a linear sign-preserving representation.

Allowing for sign-preserving transformations extends greatly the class of models covered by the maintained assumption. The following lemma gives three classes of ostensibly nonlinear functions that have linear sign-preserving representations.
Lemma. Let \([m_\gamma(\ast), \gamma \in \Gamma]\) be a collection of strictly increasing functions mapping \(R^1\) into \(R^1\). If any of the following conditions hold, \(f(x, u)\) is a sign-preserving transformation of \(x\beta + u\). (a) For each \((x, u) \in X \times R^1\), there exists \(\gamma \in \Gamma\) such that \(f(x, u) \equiv m_\gamma(x\beta + u) - m_\gamma(0)\). (b) For each \((x, u) \in X \times R^1\), there exists \(\gamma \in \Gamma\) such that \(f(x, u) \equiv -m_\gamma(-x\beta) + m_\gamma(u)\). (c) The functions \([m_\gamma(\ast), \gamma \in \Gamma]\) are antisymmetric about 0. For each \((x, u) \in X \times R^1\), there exists \(\gamma \in \Gamma\) such that \(f(x, u) \equiv m_\gamma(x\beta) + m_\gamma(u)\).

Proof. (a) The proof is immediate. (b) \(x\beta + u \geq 0 \iff u \geq -x\beta \iff m_\gamma(u) \geq m_\gamma(-x\beta) \iff -m_\gamma(-x\beta) + m_\gamma(u) \geq 0\). (c) By antisymmetry, \(-m_\gamma(-x\beta) = m_\gamma(x\beta)\).
The lemma does not exhaustively characterize the set of functions that are sign equivalent to a linear function; it does indicate that this set is rather large. The simplest applications of the lemma generate sign-preserving transformations by fixing $\gamma$ and applying the same mapping $m_\gamma$ at all $(x, u) \in X \times U$. More complex sign-preserving transformations can be obtained by letting $\gamma$ vary with $(x, u)$. 
Dual Models

Part (b) of the lemma has an intriguing consequence. It implies the existence of a class of models with a known distribution of unobservables that are dual to the linear models examined in this article.

Let $F_0$ be any continuous, strictly increasing distribution on $\mathbb{R}^1$. Let $\Gamma = X$. For each $x \in X$, let $m_x(*)$ be the strictly increasing function that transforms $F_{u|x}$ into $F_0$. That is, let $m_x(*)$ be chosen so that the distribution of $e \equiv m_x(u)$ conditional on $x$ is $F_0$. Given that $F_{u|x}$ and $F_0$ are strictly increasing, the function $m_x(*)$ exists and is unique.
By Part (b) of the lemma, \( x\beta + u \geq 0 \iff -m_x(-x\beta) + e \geq 0 \). Hence, for each \( x \in X \),

\[
P_{1|x} = \int 1[u \geq -x\beta] \, dF_{u|x} = \int 1[e \geq m_x(-x\beta)] \, dF_0.
\]

It follows that a linear latent-response model \( z = 1[x\beta + u \geq 0] \) with \( u \) distributed as \( F_{u|x} \) is equivalent to a nonlinear model \( z = 1[-m_x(-x\beta) + e \geq 0] \) with \( e \) distributed as \( F_0 \).
APPENDIX B: PROOFS OF THE COROLLARIES

Proposition 2

**Corollary 1.** This paraphrases lemma 1 of Manski (1985). The statement of that lemma assumed that $K \geq 2$, but the proof covers the case $K = 1$ as well.

**Corollary 2.** Lemma 2 of Manski (1985) proved this, given a stronger version of Condition X3. Specifically, $x_K$ was assumed to have everywhere positive Lebesgue density conditional on $w$. The proof holds without modification under the weaker assumption that $x_K$ has positive probability of being in every open interval of $R^1$, conditional on $w$. 
APPENDIX B: PROOFS OF THE COROLLARIES

Proposition 2

Corollary 1. This paraphrases lemma 1 of Manski (1985). The statement of that lemma assumed that \( K \geq 2 \), but the proof covers the case \( K = 1 \) as well.

Corollary 2. Lemma 2 of Manski (1985) proved this, given a stronger version of Condition X3. Specifically, \( x_K \) was assumed to have everywhere positive Lebesgue density conditional on \( w \). The proof holds without modification under the weaker assumption that \( x_K \) has positive probability of being in every open interval of \( R^1 \), conditional on \( w \).
Corollary 3.  (a) It suffices to consider the case $\beta_K > 0$. The case $\beta_K < 0$ is symmetric. By Proposition 2, $\text{sgn}(\beta_K)$ is identified if $\Pr(x \in Q_b) > 0$ for all $b \in R^K$ such that $b_K \leq 0$. Let $\gamma = (\beta_1, \ldots, \beta_{K-1})$ and $c = (b_1, \ldots, b_{K-1})$. Fix $w \in W$. For each $b$ such that $b_K < 0$,

\[ \Pr(x \in Q_b \mid w) \]

\[ = \Pr(xb < 0 \leq x\beta \mid w) + \Pr(x\beta < 0 \leq xb \mid w) \]

\[ = \Pr(b_Kx_K + wc < 0 \leq \beta_Kx_K + w\gamma \mid w) \]

\[ + \Pr(\beta_Kx_K + w\gamma < 0 \leq b_Kx_K + wc \mid w) \]

\[ \geq \Pr(b_Kx_K + wc < 0 \leq \beta_Kx_K + w\gamma \mid w) \]

\[ + \Pr(\beta_Kx_K + w\gamma < 0 \leq b_Kx_K + wc \mid w) \]

\[ = \Pr[x_K > \max(-wc/b_K, -w\gamma/\beta_K) \mid w] \]

\[ + \Pr[x_K < \min(-wc/b_K, -w\gamma/\beta_K) \mid w] \]

\[ \geq \min[\Pr(x_K > -w\gamma/\beta_K \mid w), \Pr(x_K < -w\gamma/\beta_K \mid w)] \]

\[ = \min[\Pr(x\beta > 0 \mid w), \Pr(x\beta < 0 \mid w)]. \]
Hence $w \in W_\beta \Rightarrow \Pr(x \in Q_b \mid w) > 0$. For each $b$ such that $b_K = 0$,

$$\Pr(x \in Q_b \mid w) = \Pr(wc < 0 \leq x\beta \mid w) + \Pr(x\beta < 0 \leq wc \mid w)$$

$$= \min[\Pr(x\beta \geq 0 \mid w), \Pr(x\beta < 0 \mid w)].$$

So again, $w \in W_\beta \Rightarrow \Pr(x \in Q_b \mid w) > 0$. It follows that for all $b$ such that $b_K \leq 0$, $\Pr(w \in W_\beta) > 0 \Rightarrow \Pr(x \in Q_b) > 0$. 
(b) By Proposition 2, $\beta_K$ is identified if $\Pr(x \in Q_b) > 0$ for all $b \in R^K$ such that $b_K \neq 0$. Fix $w \in W$. For each $b$ such that $b_K \neq 0$,

$$\Pr(x \in Q_b \mid w) = \Pr(xb < 0 \leq w \gamma \mid w) + \Pr(w \gamma < 0 \leq xb \mid w)$$

$$= \min[\Pr(xb < 0 \mid w), \Pr(xb \geq 0 \mid w)].$$

So $w \in W_b \Rightarrow \Pr(x \in Q_b \mid w) > 0$. Hence $\Pr(w \in W_b) > 0 \Rightarrow \Pr(x \in Q_b) > 0$. It follows that $\beta_K$ is identified if $\Pr(w \in W_b) > 0$ for all $b$ with $b_K \neq 0$. 
Corollary 4. Conditional on $x$, $z = 1$ with probability $P_{1|x}$ and $z = 0$ otherwise. Hence, for any $p \in [0, 1]$,

$$E[L_\alpha(p, z) \mid x] = \alpha(1 - p)P_{1|x} + (1 - \alpha)p(1 - P_{1|x})$$

$$= \alpha P_{1|x} + p[1 - \alpha - P_{1|x}].$$

It follows that $E[L_\alpha(p, z) \mid x]$ is minimized at $p = 1$ if $1 - \alpha - P_{1|x} \leq 0$ and at $p = 0$ if $1 - \alpha - P_{1|x} \geq 0$. It was shown in (6) that under Assumption QI, $x\beta \geq 0 \iff P_{1|x} \geq 1 - \alpha$. Given this, (9) follows from (8).
Corollary 5. By Assumption SI, \( u \) is quantile independent of \( x \) for every quantile \( \alpha \). It follows, by Corollary 2 (Prop. 2), that 
\[
(\beta_1 + u_\alpha, \beta_2, \ldots, \beta_K)/\|\beta_1 + u_\alpha, \beta_2, \ldots, \beta_K\| \text{ is identified for every } \alpha.
\]  
Here \( u_\alpha \) is the \( \alpha \) quantile of \( u \). \( \beta_1 + u_\alpha \) is the intercept parameter obtained when Corollary 2 is applied to quantile \( \alpha \).

The scale \( \|\beta_1 + u_\alpha, \beta_2, \ldots, \beta_K\| \) varies with \( \alpha \). But Condition X3 states that \( \beta_K \neq 0 \). It follows that the vectors \( (\beta_1 + u_\alpha, \beta_2, \ldots, \beta_K) \), \( \alpha \in (0, 1) \), can be renormalized to a common scale, namely that of \( \beta_K \). That is, 
\[
[\beta_1 + u_\alpha, \ldots, \beta_2, \ldots, \beta_K]/|\beta_K|, \alpha \in (0, 1), \text{ is identified.}
\]

To normalize location, Assumption IS sets \( u_5 = 0 \). This identifies \( \beta_1/|\beta_K| \) and \( u_\alpha/|\beta_K|, \alpha \in (0, 1) \). So \( \beta \) and \( [u_\alpha, \alpha \in (0, 1)] \) are identified up to a common scale factor. Knowledge of \( [u_\alpha, \alpha \in (0, 1)] \) is equivalent to knowledge of \( F_0 \). Hence \( (\beta, F_0) \) is identified up to scale.
Proposition 3

Corollary 1. Without loss of generality, let $k = K$. By (17), $\partial P_{1|\xi}/\partial x_K \neq 0$ implies that $dh(\xi \beta)/dv \neq 0$ and $\beta_K \neq 0$. It follows that for each $j \in [2, \ldots, K]$,

$$[\partial P_{1|\xi}/\partial x_j]/[\partial P_{1|\xi}/\partial x_K] = \beta_j/\beta_k.$$ 

Given Condition X4 and continuity of $\partial P_{1|x}/\partial x'$ at $\xi$, knowledge of $P_{1|x}$ (a.e. $F_x$) reveals the partial derivatives $\partial P_{1|\xi}/\partial x_j$, $j \in [2, \ldots, K]$. Hence the ratios $\beta_j/\beta_K$, $j \in [2, \ldots, K]$, are identified.
Proposition 4

Corollary 1. Corollary 1 for Proposition 2 showed that \( \Pr(x \in Q_b) = 0 \) for all \( b \) such that \( |b_k - \beta_k| < \lambda_1/K\mu_1 \) \((k = 1, \ldots, K)\). Hence we need to consider \( \Pr[(x, \xi) \in R_b] \). Under Condition X5, the random variable \( x - \xi \) is multinomial with positive probability mass on the point \( 0 \in R^k \) and on each of the points \( x^m - x^n \) \((m, n = 1, \ldots, M; m \neq n)\). Observe that \( x - \xi = 0 \Rightarrow (x - \xi)b = 0 \) for all \( b \in R^k \). Hence \( x - \xi = 0 \Rightarrow (x, \xi) \notin R_b \). It follows that

\[
\Pr[(x, \xi) \in R_b] = \Pr[(x, \xi) \in R_b \mid x \neq \xi] \times \Pr(x \neq \xi).
\]

Conditional on the event \([x \neq \xi]\), the distribution of \( x - \xi \) satisfies Condition X2, with \( \lambda = \lambda_2 \) and \( \mu = \mu_2 \). So Corollary 1 (Prop. 2) implies that \( \Pr[(x, \xi) \in R_b] = 0 \) for all \( b \) such that \( |b_k - \beta_k| < \lambda_2/K\mu_2 \) \((k = 1, \ldots, K)\).
Corollary 2. Let $b \in R^k$. For $(x, \zeta) \in X \times X$, let $y = (x_2 - \zeta_2, \ldots, x_k - \zeta_k)$. Let $Y$ be the domain of $y$. Let $\gamma = (\beta_2, \ldots, \beta_k)$ and $c = (b_2, \ldots, b_k)$. Define $Y_b = \{y \in Y : yc < 0 \leq y\gamma \cup y\gamma < 0 \leq yc\}$. Recall that $x_1 = \zeta_1 = 1$. Hence $(x, \zeta) \in R_b \iff y \in Y_b$. It will be shown that $\Pr[y \in Y_b] > 0$ for all $b$ such that $c/\|c\| \neq \gamma/\|\gamma\|$. Condition X6 implies that conditional on almost every value of $(y_1, \ldots, y_{k-2})$, $y_{k-1}$ has positive probability mass in every open interval of $R^1$. Condition X1 implies that $F_y$ does not place probability 1 in any proper linear subspace of $R^{k-1}$. So the distribution of $y$ satisfies Conditions X1 and X3. The result now follows from Corollary 2 for Proposition 2.
Corollary 3. (a) Part (a) of Corollary 3 for Proposition 2 showed that \( \Pr(w \in W_{\beta}) > 0 \) identifies \( \text{sgn}(\beta_K) \). Given Proposition 4, the same proof holds here with \( d \) replacing \( w \) and \( D_{b} \) replacing \( W_{b} \). That is, either \( \Pr(w \in W_{\beta}) > 0 \) or \( \Pr(d \in D_{\beta}) > 0 \) suffices.

(b) The argument is the same as in part (a).
ESTIMABILITY IN THE MULTINOMIAL PROBIT MODEL

David S. Bunch
Graduate School of Management, University of California, Davis, Davis, CA 95616, U.S.A.

(Received 11 July 1989; in revised form 15 February 1990)

Abstract — Random utility models often involve terms which represent alternative-specific errors, and the main attractive feature of the multinomial probit (MNP) model is that it allows a rather general covariance structure for these errors. However, since observed choices only reveal information regarding utility differences, and since scale cannot be determined, not all parameters in an arbitrary MNP specification may be identified. This paper examines identification restrictions that arise in the linear-in-parameters multinomial probit framework, and provides discussion and recommendations for estimation and analysis of probit normalizations.
1. INTRODUCTION

Recent advances in computational methods for estimating multinomial probit models have stimulated renewed interest on this topic: for example, McFadden (1989) and Pakes and Pollard (1989) suggest simulation methods which may lead to practical probit estimation codes for more than four alternatives. In addition, Kamakura (1989) demonstrates, via a simulation study, that the Mendell-Elston approximation to the multivariate normal CDF is more accurate than the Clark method, and may provide an alternative solution to the problem. Bunch and Kitamura (1989) corroborate Kamakura's results in a study using empirical data, and discuss improved algorithms for maximum likelihood estimation.

This paper considers another important practical issue, the problem of formulating multinomial probit (MNP) model specifications for which the parameters are estimable. A discussion of parameter estimability issues and some illustrative examples appear in section 3.1 of Daganzo (1979); however, the topic is difficult and no general comprehensive theory is offered by Daganzo, nor do we believe one is likely to be offered in the near future. The discussion and conclusions presented here focus on the linear-in-parameters MNP model with taste variation and correlated random errors. This framework is quite flexible and is consistent with many specifications discussed in the literature.
The identification difficulties which arise are primarily due to the random errors, which are usually associated with the effect of unobserved attributes on a choice object's utility. A general model specification includes the possibility of correlations among the utilities of objects which might share unobserved attributes. Unfortunately, information about the underlying utilities is available only through observation of discrete choices, which depend on differences of utilities. This, plus the issue of scaling, lead to restrictions on the number of estimable parameters.

The material presented here overlaps that of Dansie (1985) and Albright, Lerman, and Manski (1977), and is perhaps well-known to some (certainly not all) workers in the field. However, the extent to which this is true is unclear. Bunch and Kitamura (1989) give a brief review of empirical applications, and a significant proportion of them were found to contain model specification errors. This is important since misspecified models potentially compound the already troublesome computational difficulties inherent in computing MNP estimates, and could be a contributing factor to the relative dearth of successful empirical applications of MNP in the published literature.
2. LINEAR-IN-PARAMETERS MULTINOMIAL PROBIT MODELS

In this paper we will consider a standard discrete choice modeling situation in which an individual drawn at random from a population makes a choice from a set of $J$ mutually exclusive alternatives. A common example in the transportation literature is the choice of mode for the work commute, where $J = 3$ or $4$ and the alternatives consist of car, train, bus, shared ride, etc. Following McFadden (1981, 1989), assume that the utility of alternative $j$ for individual $n$, $U_{nj}$, is given by the general form

$$U_{nj} = X_{nj} \alpha_n, j = 1, \ldots, J,$$

where $X_{nj}$ is a $K$-vector of explanatory variables which may be a function of the attributes of alternative $j$ and individual $n$. The $K$-vector $\alpha_n$ contains the taste weights for individual $n$, and may be rewritten as $\alpha_n = \theta + \delta_n$, where $\theta$ is the mean taste weight for the population and $\delta_n$ is the (unobserved) random deviation from the mean for individual $n$. (Ideally, $\theta$ and the distribution of $\delta_n$ would vary as an explicit function of the observed characteristics of the individual, but this consideration is usually suppressed for simplicity.)
In theory, utility should be a function of generic (or "real") attributes, and should not depend on nominal attributes such as the labels "car," "bus," etc. In practical applications it is also desirable for the model to include only generic variables, since the model may then be used more effectively in forecasting, especially for testing the effect of adding new choice alternatives. However, it has generally been observed that including alternative-specific dummy variables significantly improves the fit of discrete choice models. One interpretation is that unobserved attributes are often empirically correlated with the nominal labels of the alternatives (McFadden et al., 1977). For example, the attributes "lack of flexibility," or "lack of comfort," might be correlated with the nominal label "bus."
Equation (1) readily accommodates the inclusion of dummy variables, which is the approach taken by Albright, Lerman, and Manski (1977). Alternatively, we may assume that $X_{nj}$ in eqn (1) includes only generic variables, and add the additional terms $\mu_j$ and $\epsilon_{nj}$, where $\mu_j$ is the mean of the alternative-specific errors, and $\epsilon_{nj}$ represents a random deviation from the mean. The $\epsilon_{nj}$ term may be regarded to include both the effects of unobserved attributes and any other sources of observation-specific random error. This gives the following model, expressed now in vector notation:

$$U_n = X_n^T(\theta + \delta_n) + \mu + \epsilon_n,$$

(2)

where $U_n$, $\mu$, $\epsilon_n \in \mathbb{R}^J$, $\theta$, $\delta_n \in \mathbb{R}^K$, and $X_n \in \mathbb{R}^{K \times J}$. To get a multinomial probit model, one adds the theoretically appealing assumption that the random terms have multivariate normal distributions:

$$\delta_n \sim MVN(0, \Sigma_\delta), \Sigma_\delta \in \mathbb{R}^{K \times K} \text{ and } \epsilon_n \sim MVN(0, \Sigma_\epsilon), \Sigma_\epsilon \in \mathbb{R}^{J \times J}.$$

(3)

Note that in this formulation the $\delta$ and $\epsilon$ terms are assumed to be independent, which is slightly more restrictive that the model implied by eqn (1). This framework is consistent with Hausman and Wise (1978).
Now, the probability that individual $n$ selects alternative $j$ is given by the MNP model:

$$P(j \mid V_U(\theta, \mu, X_n), \Sigma_U(\Sigma_\delta, \Sigma_\epsilon, X_n)) = \text{Prob}[U_{nj} > U_{ni} \text{ for all } i \neq j] \quad (4a)$$

$$= \int_{u_j = -\infty}^{\infty} \ldots \int_{u_j = -\infty}^{u_j} \int_{u_{j-1} = -\infty}^{u_{j-1}} \ldots \int_{u_1 = -\infty}^{u_1} \phi(u \mid V_U, \Sigma_U) du_1 \cdots du_j \quad (4b)$$

where

$$V_U(\theta, \mu, X_n) = X_n^T \theta + \mu,$$

$$\Sigma_U(\Sigma_\delta, \Sigma_\epsilon, X_n) = X_n^T \Sigma_\delta X_n + \Sigma_\epsilon,$$

and $\phi(x \mid m, S)$ is the multivariate normal density function with mean $m$ and covariance $S$. If we assume that $\Sigma_\delta$ and $\Sigma_\epsilon$ are positive definite then it is straightforward to show that $\Sigma_U(\Sigma_\delta, \Sigma_\epsilon, X_n)$ is also positive definite, which is desirable for establishing regularity conditions: see Daganzo (1979), or, hereafter, “Daganzo.”
The purpose of this paper is to discuss the issues and problems of specifying the linear-in-parameters MNP model so that all the parameters are estimable. A primary concern is specification of $\Sigma_r$, which is especially important since most published applications of MNP focus exclusively on estimating the effects of observation-specific errors (i.e. they assume (2) with $\delta_n = 0$).
3. REDUCTION OF DIMENSION

As has been often noted, one of the difficulties with MNP is that it requires evaluation of the multivariate integral (4b), which does not have a closed form solution. The usual first step is to reduce the dimension of the integral from \( J \) to \( J-1 \) using the transformation discussed by Daganzo (1979, pp. 43–44 and pp. 94–95). Assume that we wish to compute \( P(j \mid V_u(\theta, \mu, X), \Sigma_u(\Sigma_\theta, \Sigma_c, X)) \) and define \( \Delta_j \in \mathbb{R}^{J-1 \times J} \) by

\[
\Delta_j = \begin{bmatrix}
1 & 2 & 3 & \cdots & j-1 & j & j+1 & \cdots & J-1 & J \\
1 & 0 & 0 & \cdots & 0 & -1 & 0 & \cdots & 0 & 0 \\
2 & 0 & 1 & 0 & \cdots & 0 & -1 & 0 & \cdots & 0 & 0 \\
3 & 0 & 0 & 1 & \cdots & 0 & -1 & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
j-1 & 0 & 0 & 0 & \cdots & 1 & -1 & 0 & \cdots & 0 & 0 \\
j & 0 & 0 & 0 & \cdots & 0 & -1 & 1 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
J-2 & 0 & 0 & 0 & \cdots & 0 & -1 & 0 & \cdots & 1 & 0 \\
J-1 & 0 & 0 & 0 & \cdots & 0 & -1 & 0 & \cdots & 0 & 1 
\end{bmatrix}.
\]
The transformation $Z = \Delta_j U$ applied to (4) gives

$$P(j \mid V_U(\theta, \mu, X), \Sigma_U(\Sigma_\delta, \Sigma_\epsilon, X)) = \text{Prob} [U_i - U_j < 0, i \neq j]$$

$$= \Phi(0 \mid V_Z, \Sigma_Z)$$

where

$$V_Z = \Delta_j V_U = \Delta_j (X^T\theta + \mu) = \Delta_j X^T\theta + \Delta_j \mu,$$

$$\Sigma_Z = \Delta_j \Sigma_U \Delta_j^T = \Delta_j (X^T \Sigma_\delta X + \Sigma_\epsilon) \Delta_j^T = \Delta_j X^T \Sigma_\delta X \Delta_j^T + \Delta_j \Sigma_\epsilon \Delta_j^T;$$

$Z, V_Z \in \mathbb{R}^{J-1}, \Sigma_Z \in \mathbb{R}^{(J-1) \times (J-1)}$, and $\Phi$ is the cumulative distribution function for the $(J - 1)$-dimensional multivariate normal distribution.
Define \( m_j \in \mathcal{R}^{J-1} \) by \( m_j = \Delta_j \mu \) and \( C_j \in \mathcal{R}^{(J-1) \times (J-1)} \) by \( C_j = \Delta_j \Sigma_\epsilon \Delta_j^T \), where \( C_j \) is symmetric. Consider the matrix \( M_j \in \mathcal{R}^{(J-1) \times (J-1)} \) which is obtained by taking \( \Delta_j \) and deleting the \( J \)th column; \( M_j \) is of full rank (i.e. rank \( J - 1 \)). Next, without loss of generality choose alternative \( J \) as the "reference alternative." It is straightforward to show that \( \Delta_j = M_j \Delta_j \) for all \( j \), and hence \( m_j = M_j m_j \) and \( C_j = M_j C_j M_j^T \) for all \( j \). It follows that (7) may be rewritten as

\[
V_z = \Delta_j X^T \theta + M_j m_j, \\
\Sigma_z = \Delta_j X^T \Sigma_\epsilon X \Delta_j^T + M_j C_j M_j^T,
\]

and thus choice probabilities evaluated via (6) may always be expressed in terms of \( m_j \) and \( C_j \), which together contain \( (J - 1) + J(J - 1)/2 \) parameters. It follows that \( \mu \) and \( \Sigma_\epsilon \) together only have \( (J - 1) + J(J - 1)/2 \) identifiable parameters. As an illustration, consider the case \( J = 3 \) and \( j = 2 \), which gives (taking into account the symmetry of \( \Sigma_\epsilon \)): 
\[
m_2 = \begin{bmatrix}
\mu_1 - \mu_2 \\
\mu_3 - \mu_2
\end{bmatrix} = \begin{bmatrix}
1 & -1 \\
0 & -1
\end{bmatrix} \begin{bmatrix}
\mu_1 - \mu_3 \\
\mu_2 - \mu_3
\end{bmatrix} = M_2 m_3,
\]

and

\[
C_2 = \begin{bmatrix}
\sigma_{11} - 2\sigma_{21} + \sigma_{22} & \sigma_{21} - \sigma_{31} - \sigma_{32} + \sigma_{22} \\
\sigma_{21} - \sigma_{31} - \sigma_{32} + \sigma_{22} & \sigma_{33} - 2\sigma_{31} + \sigma_{22}
\end{bmatrix}
\]

\[
= \begin{bmatrix}
1 & -1 \\
0 & -1
\end{bmatrix} \begin{bmatrix}
\sigma_{11} - 2\sigma_{31} + \sigma_{33} & \sigma_{21} - \sigma_{31} - \sigma_{32} + \sigma_{33} \\
\sigma_{21} - \sigma_{31} - \sigma_{32} + \sigma_{33} & \sigma_{22} - 2\sigma_{32} + \sigma_{33}
\end{bmatrix} \begin{bmatrix}
1 & 0 \\
-1 & -1
\end{bmatrix}
\]

\[
= M_2 C_3 M_2^T,
\]

where the \( \sigma_{ij} \)'s denote the elements of \( \Sigma_\epsilon \).
In addition to these considerations, it turns out that one can also rescale the problem so as to eliminate another parameter. Specifically, evaluation of $\Phi(0 \mid kV_z, k^2\Sigma_z)$ gives the same result as evaluating $\Phi(0 \mid V_z, \Sigma_z)$, for $k > 0$. Hence, suitable selection of $k$ can eliminate one more parameter. This leads to the assertion that the model (3) has a total of $K + K(K + 1)/2 + J + J(J - 1)/2 - 2$ identifiable parameters. Since $\theta$, $\mu$, $\Sigma$, and $\Sigma_e$ together have $K + K(K + 1)/2 + J + J(J + 1)/2$ parameters, this implies that $J + 2$ parameters are inestimable and must be fixed via some normalization. (The fact that $\Sigma_e$ has only $J(J - 1)/2$ identifiable parameters was noted by Daganzo, who also recognized that one could sometimes eliminate another parameter by rescaling. Albright, Lerman, and Manski (1977) state the above result for the number of free parameters, referring the reader to technical memoranda.)
Note that for the binary probit model with no taste variation (i.e. \( J = 2 \) and \( \delta = 0 \)),
there are \( K \) taste weights, \( J - 1 \) alternative specific dummy variables, and \( J(J - 1)/2 - 1 = 0 \) free parameters in \( \Sigma_z \). Although this is generally known, even experienced investigators publishing in refereed journals can miss this point; for example, in a study of travel mode choice Johnson and Hensher (1982) erroneously assume that \( \Sigma_z \) has one free parameter. In studies involving more complex MNP models, other authors have completely missed the requirements described above, and assume that all the parameters are in principle estimable. See, for example, Currim (1982) and van Lierop (1986). For a brief review of MNP empirical applications which includes further discussion, see Bunch and Kitamura (1989).
These restrictions have implications for practical estimation of MNP models. Recall that Daganzo discusses MNP in a slightly more general framework by considering the MNP function \( P(j \mid V(\theta, A), \Sigma(\theta, A)) \), where \( \theta \) is a general parameter vector and \( A \) is a matrix of attributes. (Note, however, that many of his examples fit in the linear-in-parameters framework.) In discussing the practical issues of parameter estimation, he recommends that \( \Sigma(\theta, A) \) be specified so as to be positive definite over all feasible values of \( \theta \), since otherwise it “would not represent a covariance matrix and the program would not return meaningful values.” Daganzo recommends two possibilities:

(i) express \( \Sigma(\theta, A) \) as a product of a matrix and its transpose (i.e. \( \Sigma(\theta, A) = C(\theta, A)C(\theta, A)^T \)),

(ii) express \( \Sigma(\theta, A) \) as a function of \( \theta \) and \( A \) directly, placing simple bounds on \( \theta \) so as to ensure positive definiteness.
We may regard $C(\theta, A)$ to be lower triangular with positive diagonal elements (i.e. the Cholesky factorization of $\Sigma(\theta, A)$). For the case we are considering here, either of these approaches could be difficult when choosing a specification for $\Sigma_U$. Directly writing $\Sigma_U = CC^T$ is not practical if taste variation is included in the model. One could express each of $\Sigma_6$ and $\Sigma_\varepsilon$ in terms of Cholesky factorizations, but this could be tricky for $\Sigma_\varepsilon$ (or, $\Sigma_U$ in the case of fixed tastes) since we have at most $J(J - 1)/2$ free parameters to work with in $\Sigma_\varepsilon$. For most of the following discussion, we will assume that we wish to choose a normalization which utilizes all available parameters.
Albright, Lerman, and Manski (1977), which we denote “ALM” in the sequel, deal with the issue by choosing one normalization from “among the alternative formally equivalent normalizations.” First, they choose to express both $\Sigma_\delta$ and $\Sigma_\epsilon$ in terms of Cholesky factorizations. Next, they choose to fix $J$ parameters by setting the last row (and column) of $\Sigma_\epsilon$, and correspondingly, the last row of its Cholesky factorization, to zero. Finally, they choose to fix the scale of the specification by constraining the diagonal elements of $\Sigma_\epsilon$ so that $(\text{trace } \Sigma_\epsilon)/J$ equals the variance of the standard Weibull distribution. (This was done to facilitate comparisons between MNP and multinomial logit estimates.) Note that this is inconsistent with Daganzo’s recommendation in the no-taste-variation case: this specification of $\Sigma_\epsilon$, while a valid covariance matrix, is only positive semi-definite and hence violates Daganzo’s regularity conditions.
In fact, ALM are doing slightly more than simply choosing an "arbitrary normalization": they are choosing to work directly in \((J - 1)\)-space, estimating \(C_j\) so that \(C_j\) is constrained to be positive definite (which implies that \(C_j\) for all \(j \neq J\) must also be positive definite). This is a more general approach than choosing to perform the estimation using arbitrarily constrained formulations in the original \(J\)-dimensional space.

As an example, consider the case \(J = 3\). For simplicity, we suppress the scaling issue at this stage: one can assume that scale has been fixed via elimination of a parameter component from either \(\theta\), \(\mu\), or \(\Sigma_3\). For \(J = 3\) \(\Sigma_\epsilon\) has \(3(3 - 1)/2 = 3\) free parameters. The ALM normalization (with no rescaling) is given by:

\[
\Sigma_\epsilon^{ALM} = \begin{bmatrix}
\sigma_{11} & \sigma_{21} & 0 \\
\sigma_{21} & \sigma_{22} & 0 \\
0 & 0 & 0
\end{bmatrix}.
\]
Another possible normalization, which assumes independence of the random errors, is given by a diagonal covariance matrix:

$$\Sigma^D = \begin{bmatrix} \sigma_{11} & 0 & 0 \\ 0 & \sigma_{22} & 0 \\ 0 & 0 & \sigma_{33} \end{bmatrix},$$

where \( \sigma_{11}, \sigma_{22}, \sigma_{33} > 0 \) to ensure positive definiteness. Now, suppose that a maximum likelihood estimation routine produces an estimate for \( C_j \) and returns the following result:

$$C_j = \begin{bmatrix} 1.5 & -0.4 \\ -0.4 & 1.2 \end{bmatrix}$$

Then the ALM normalization is simply

$$\Sigma^\text{ALM} = \begin{bmatrix} 1.5 & -0.4 & 0 \\ -0.4 & 1.2 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$
However, the corresponding $\Sigma^D$ is computed to be:

\[
\Sigma^D = \begin{bmatrix} 1.9 & 0 & 0 \\ 0 & 1.6 & 0 \\ 0 & 0 & -0.4 \end{bmatrix},
\]

which is clearly not a valid covariance matrix. Problems could arise if the estimation routine were attempting to directly estimate the diagonal normalization: a boundary solution would probably be the result.

It seems clear that the best course to follow is to first estimate $C_j$, and to then explore the various possible normalizations by performing transformations between $C_j$ and $\Sigma_e$. If any particular specification is inappropriate, it will show up as an invalid covariance matrix without jeopardizing the performance of the estimation routine. In addition, the possibility of accidentally choosing an inestimable normalization will be minimized, since the process of analyzing the transformation will reveal the mistake. This is discussed next.
4. TRANSFORMATIONS

As above, assume for now that fixing the scale is not an issue. In addition, assume that the specifications we are considering are fairly simple, (i.e. parameters do not appear simultaneously across $\theta$, $\mu$, $\Sigma_\delta$, $\Sigma_\epsilon$). (Although this restriction is not necessary, it greatly simplifies the analysis and discussion.) In what follows, it will often be convenient to regard symmetric matrices as vectors in packed storage form (e.g. when $J = 3$ then $\Sigma_\epsilon \in \mathbb{R}^{J(J+1)/2}$ with $\Sigma_\epsilon = (\sigma_{11}, \sigma_{21}, \sigma_{22}, \sigma_{31}, \sigma_{32}, \sigma_{33})^T$, and $C_J \in \mathbb{R}^{(J-1)/2}$ with $C_J = (c_{11}, c_{21}, c_{22})^T$).

Suppose that $\beta \in \mathbb{R}^{(J-1)/2}$ denotes the vector of "identified" parameters. Then candidate normalizations may be represented as mappings from $\mathbb{R}^{(J-1)/2}$ to $\mathbb{R}^{J(J+1)/2}$. For example, consider the ($J = 3$) ALM normalization in eqn (9) which in matrix form can be written as

$$
\Sigma_\epsilon^{\text{ALM}} = \begin{bmatrix}
\beta_1 & \beta_2 & 0 \\
\beta_2 & \beta_3 & 0 \\
0 & 0 & 0
\end{bmatrix}.
$$

(14)
In packed form this particular normalization may be represented by the (linear) mapping

\[
\begin{bmatrix}
\beta_1 \\
\beta_2 \\
\beta_3 \\
0 \\
0 \\
0
\end{bmatrix}
= h_1(\beta) =
\begin{bmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1 \\
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
\beta_1 \\
\beta_2 \\
\beta_3
\end{bmatrix}
= H_1 \beta.
\] (15)

Recall that the “observable” parameters reside in \( C_j \in \mathbb{R}^{(J-1)J/2} \) (in packed form), which is obtained through the mapping \( g(\cdot) \) from \( \mathbb{R}^{J(J+1)/2} \) to \( \mathbb{R}^{(J-1)J/2} \), defined below. Define the function \( f(\beta) = g(h(\beta)) \) which maps \( \mathbb{R}^{(J-1)J/2} \) to itself. Then the identification problem reduces to verifying that \( f(\cdot) \) represents a unique invertible transformation between \( \beta \) and \( C_j \). The relevant requirement from the Inverse Function Theorem—see Apostol (1974)—is that the Jacobian determinant of \( f \) be different from zero.
To illustrate, consider the simple class of $J = 3$) normalizations which are restricted so that either $\sigma_{ij} = 0$ or $\sigma_{ij} = \sigma_{kl}$. This includes both of the previous examples:

$$\Sigma_1 = \begin{bmatrix} \beta_1 & \beta_2 & 0 \\ \beta_2 & \beta_3 & 0 \\ 0 & 0 & 0 \end{bmatrix} \Rightarrow h_1(\beta) = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \beta = H_1\beta$$ (16)

and

$$\Sigma_2 = \begin{bmatrix} \beta_1 & 0 & 0 \\ 0 & \beta_2 & 0 \\ 0 & 0 & \beta_3 \end{bmatrix} \Rightarrow h_2(\beta) = \begin{bmatrix} \beta_1 \\ 0 \\ \beta_2 \\ 0 \\ 0 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \beta = H_2\beta.$$ (17)
Some additional possibilities are:

\[
\Sigma_3 = \begin{bmatrix} \beta_1 & \beta_2 & 0 \\ \beta_2 & \beta_1 & \beta_3 \\ 0 & \beta_3 & \beta_1 \end{bmatrix} \Rightarrow h_3(\beta) = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_1 \\ 0 \\ \beta_3 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \beta = H_3\beta, \quad (18)
\]

\[
\Sigma_4 = \begin{bmatrix} \beta_1 & \beta_2 & 0 \\ \beta_2 & \beta_3 & 0 \\ 0 & 0 & \beta_3 \end{bmatrix} \Rightarrow h_4(\beta) = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ 0 \\ 0 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \beta = H_4\beta, \quad (19)
\]

\[
\Sigma_5 = \begin{bmatrix} \beta_1 & \beta_2 & 0 \\ \beta_2 & \beta_1 & 0 \\ 0 & 0 & \beta_3 \end{bmatrix} \Rightarrow h_5(\beta) = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_1 \\ 0 \\ 0 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \beta = H_5\beta, \quad (20)
\]

\[
\Sigma_6 = \begin{bmatrix} \beta_1 & \beta_2 & \beta_3 \\ \beta_2 & \beta_1 & \beta_3 \\ \beta_3 & \beta_3 & \beta_1 \end{bmatrix} \Rightarrow h_6(\beta) = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_1 \\ \beta_3 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \beta = H_6\beta. \quad (21)
\]
The function $g(\cdot)$ which maps $\Sigma$ to $C_j$ may be defined by the linear transformation $G$ given by

\[ g(\Sigma) = \begin{bmatrix} 1 & 0 & 0 & -2 & 0 & 1 \\ 0 & 1 & 0 & -1 & -1 & 1 \\ 0 & 0 & 1 & 0 & -2 & 1 \end{bmatrix} \begin{bmatrix} \sigma_{11} \\ \sigma_{21} \\ \sigma_{22} \\ \sigma_{31} \\ \sigma_{32} \\ \sigma_{33} \end{bmatrix} = G\Sigma, \]  

and hence for this class of normalizations $f(\beta) = g(h(\beta)) = GH\beta$ is the composite of two linear transformations. Application of the chain rule gives the Jacobian of $f$, $Df = Dg \circ Dh = GH$. It follows that

\[
Df_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad Df_2 = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}, \\
Df_3 = \begin{bmatrix} 2 & 0 & 0 \\ 2 & 0 & -2 \\ 1 & 1 & -1 \end{bmatrix}, \quad Df_4 = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 2 \end{bmatrix}, \\
Df_5 = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}, \quad Df_6 = \begin{bmatrix} 2 & 0 & -2 \\ 1 & 1 & -2 \\ 2 & 0 & -2 \end{bmatrix}.
\]
The Jacobians $Df_1$ through $Df_4$ are of full rank, and hence the $\beta$'s are identifiable for the corresponding specifications. In contrast, $Df_5$ and $Df_6$ are not of full rank, and the $\beta$'s are not identifiable in $\Sigma_5$ and $\Sigma_6$. Thus, one may not arbitrarily choose any normalization which appears to have the correct number of free parameters. It is also essential to verify that the transformation implied by the normalization is invertible, a point which has been overlooked until recently.

In particular, consider $\Sigma_5$. If the scaling of this specification is fixed by dividing by $\beta_1$, the resulting normalization is

$$
\Sigma'_5 = \begin{bmatrix}
1 & \beta'_2 & 0 \\
\beta'_2 & 1 & 0 \\
0 & 0 & \beta'_3
\end{bmatrix}.
$$

This unidentified normalization appears in Daganzo (1979), is quoted by Currim (1982), and is used in a simulation study by Horowitz et al. (1982). The fact that $\Sigma'_5$ is unidentified was pointed out by Dansie (1985).
The above approach is general and is easily extended to $J > 3$. For example, consider a mode choice problem with $J = 4$ where the choice set is \{drive alone, carpool, train, bus\}. There are $4(4 - 1)/2 = 6$ identifiable parameters in $\Sigma_e$, and two possible candidate normalizations are:

\[
\Sigma_6 = \begin{bmatrix}
\beta_1 & \beta_2 & 0 & 0 \\
\beta_2 & \beta_3 & 0 & 0 \\
0 & 0 & \beta_4 & \beta_5 \\
0 & 0 & \beta_5 & \beta_6
\end{bmatrix},
\]

and

\[
\Sigma_7 = \begin{bmatrix}
\beta_1 & \beta_2 & \beta_3 & \beta_3 \\
\beta_2 & \beta_3 & \beta_4 & \beta_4 \\
\beta_3 & \beta_4 & \beta_5 & \beta_6 \\
\beta_3 & \beta_4 & \beta_6 & \beta_5
\end{bmatrix}.
\]

For $\Sigma_6$, the assumption is that each of the alternatives has an error term with a different variance, and that there is correlation of unobserved attributes between the pairs \{drive alone, carpool\} and \{train, bus\}, respectively. Error terms for \{drive, carpool\} are assumed to be uncorrelated with error terms for \{train, bus\}. In contrast, for $\Sigma_7$ the error terms for \{drive, carpool\} are assumed to be correlated with the error terms for \{train, bus\}, and each of these pairs is assumed to have the same variance. A straightforward application of the procedure developed above reveals that $\Sigma_6$ is identified, while $\Sigma_7$ is not.
5. SCALING

Although one can generally consider fixing the scale of (2) by dividing through by a coefficient parameter, the standard practice in published MNP applications is to incorporate scaling in the specification of $\Sigma_e$. This seems more intuitive, and is based on what are probably more acceptable assumptions. The identification results for the examples of the previous section still hold if one chooses to include scaling, as a consequence of the simple structure that was assumed.

If thinking about the specification in $J$-space, one could divide $\Sigma_e$ by a constant multiple of one of the (nonzero) variance components, leaving one or more constant terms on the diagonal. (If one is keeping track, a corresponding adjustment would take place for $\theta$ and $\mu$, multiplying them by the positive square root of the same expression.) Alternatively, one could consider the “observable” parameters to live in $(J - 1)$-space, and divide all the elements of $C_J$ by $c_{11} (= \sigma_{11} - 2\sigma_{1J} + \sigma_{JJ})$ so that the estimated parameters reside in a symmetric $(J - 1)$-dimensional matrix with a “1” in the upper left-hand corner. The matrix $C_J$ may be constrained to be positive definite through a convenient Cholesky factorization $B_J$ with $J(J - 1)/2 - 1$ free parameters (i.e. $C_J = B_J B_J^T$) where:

$$B_J = \begin{bmatrix}
1 & 0 & 0 & \cdots & 0 \\
b_{21} & b_{22} & 0 & \cdots & 0 \\
b_{31} & b_{32} & b_{33} & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
b_{J1} & b_{J2} & b_{J3} & \cdots & b_{JJ}
\end{bmatrix}.$$ 

(26)
Transformations which directly incorporate scaling, defined using the spaces $\mathcal{H}_{J(\mathcal{J}+1)/2-1}$ and $\mathcal{H}_{(J-1)/2-1}$, may be studied using the ideas of the previous section, although the exercise is more tedious and a bit less intuitive. It should be clear that one can readily move from one normalization to another via $C_J$, and rescale when it is convenient. This is illustrated in the next section.
6. NUMERICAL EXAMPLES

A major application of choice models in transportation is the estimation of commuter mode choice. Most applications of this type have used the logit model, but two major examples using linear-in-parameters MNP are Hausman and Wise (1978) and the report by Albright, Lerman, and Manski. Both use a data set collected in Washington, D.C. in which the choices are \{car, shared-ride, bus\}.

Hausman and Wise (HW) use a subset containing 100 observations, and restrict themselves to models for which the most general specification assumes uncorrelated random errors (i.e. that $\Sigma_\delta$ and $\Sigma_\epsilon$ are diagonal). The generic attributes are trip cost divided by personal income, in-vehicle travel time, and out-of-vehicle travel time. The choice probabilities are evaluated using numerical integration.

In contrast, ALM use 1353 observations, and estimate a quite general specification which allows for correlated random errors. They use essentially the same generic attributes as HW, but include two mode-specific variables on available autos per licensed driver. They evaluate choice probabilities via Clark's approximation—see Daganzo (1979).
An interesting feature of these reported results is that both obtain estimates for a fully-specified $\Sigma_\epsilon$ matrix. In fact, the two specifications are essentially those given by eqns (9) and (10) above, but with the following adjustments for scaling purposes:

$$\Sigma^{\text{ALM}}_\epsilon = \begin{bmatrix} \sigma_{11} & \sigma_{21} & 0 \\ \sigma_{21} & w - \sigma_{11} & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

$$\Sigma^{\text{HW}}_\epsilon = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \sigma_{22} & 0 \\ 0 & 0 & \sigma_{33} \end{bmatrix}.$$
Based on the discussion in the previous two sections, one can transform the HW and ALM results into various identifiable normalizations for comparison purposes. Table 1 includes such a comparison using the (identified) normalizations discussed above. There are no striking differences between the two sets of results, even though they were obtained with different sample sizes and different assumptions regarding taste variation. It is clear from this example that, depending upon the modeling assumptions one is willing to make, different interpretations involving the effect of unobserved variables are possible.

However, recall from the example in eqns (11) through (13) that in some situations candidate normalizations could be rejected if they produce invalid covariance matrices. This is illustrated in Bunch and Kitamura (1989), in which trinomial probit models of car ownership are estimated using the approach recommended here. A comparison of various normalizations is made, with some being rejected as invalid.
7. MORE PARSIMONIOUS SPECIFICATIONS

In the section on transformations we limited ourselves to normalizations which utilize all available parameters; however, this is not necessary and one might wish to consider more parsimonious specifications involving fewer parameters. In this case the relationships between $C_j$ and $\Sigma_i$ may still be examined, but the researcher must give careful thought to what is being assumed.

For example, the specification $\Sigma_5$ (or, equivalently, $\Sigma_5'$) is unidentified. Dansie (1985) shows that the three covariance matrices

$$
\Sigma_A = \begin{bmatrix}
1 & \sigma_{21} & 0 \\
\sigma_{21} & 1 & 0 \\
0 & 0 & \sigma_{33}
\end{bmatrix}
$$

$$
\Sigma_B = \begin{bmatrix}
1 & \sigma_{21} & 0 \\
\sigma_{21} & 1 & 0 \\
0 & 0 & 1
\end{bmatrix}
$$

$$
\Sigma_C = \begin{bmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & \sigma_{33}
\end{bmatrix}
$$

are equivalent, but that only the second two are identified, with one free parameter.
As a consequence, he recommends estimating the model using a one-parameter $C_j$ matrix given by

$$C_j = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$  

This is equivalent to assuming that, whatever the normalization may look like, we expect that the variance of $\epsilon_1-\epsilon_3$ equals the variance of $\epsilon_2-\epsilon_3$. Note that many normalizations may satisfy this requirement. In particular, if the estimation routine produces a result with $\rho < 0$, then neither of the two normalizations $\Sigma_B$ and $\Sigma_C$ will be valid covariance matrices; of course others could be found which are.
Table 1. Alternative $\Sigma_c$ normalization for MNP mode choice results

<table>
<thead>
<tr>
<th>Albright, Lerman, and Manski</th>
<th>Hausman and Wise</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\begin{bmatrix} 1 &amp; 0.45 &amp; 0 \ 0.45 &amp; 1.21 &amp; 0 \ 0 &amp; 0 &amp; 0 \end{bmatrix}$</td>
<td>$\begin{bmatrix} 1 &amp; 0.50 &amp; 0 \ 0.50 &amp; 1.51 &amp; 0 \ 0 &amp; 0 &amp; 0 \end{bmatrix}$</td>
</tr>
<tr>
<td>$\begin{bmatrix} 1 &amp; 0 &amp; 0 \ 0 &amp; 1.38 &amp; 0 \ 0 &amp; 0 &amp; 0.82 \end{bmatrix}$</td>
<td>$\begin{bmatrix} 1 &amp; 0 &amp; 0 \ 0 &amp; 2.02 &amp; 0 \ 0 &amp; 0 &amp; 1.00 \end{bmatrix}$</td>
</tr>
<tr>
<td>$\begin{bmatrix} 1 &amp; -0.21 &amp; 0 \ -0.21 &amp; 1 &amp; -0.34 \ 0 &amp; -0.34 &amp; 1 \end{bmatrix}$</td>
<td>$\begin{bmatrix} 1 &amp; -0.51 &amp; 0 \ -0.51 &amp; 1 &amp; -0.51 \ 0 &amp; -0.51 &amp; 1 \end{bmatrix}$</td>
</tr>
<tr>
<td>$\begin{bmatrix} 1 &amp; -0.39 &amp; 0 \ -0.39 &amp; 1.53 &amp; 0 \ 0 &amp; 0 &amp; 1.53 \end{bmatrix}$</td>
<td>$\begin{bmatrix} 1 &amp; -1.04 &amp; 0 \ -1.04 &amp; 3.08 &amp; 0 \ 0 &amp; 0 &amp; 3.08 \end{bmatrix}$</td>
</tr>
</tbody>
</table>
8. SUMMARY

The linear-in-parameters multinomial probit framework in eqn (2) includes a random error vector which in general has a covariance matrix \( \Sigma_e \), but unfortunately not all the covariance parameters are identified. Section 4 develops the arguments which give the number of estimable parameters, and recommends that estimation be performed in terms of a matrix \( C_f \). Multiple normalizations may correspond to the same estimated \( C_f \), but some may not produce valid covariance matrices. Choosing a particular (valid) normalization is essentially a \textit{modeling decision}, as illustrated via the examples in section 4. This feature of MNP is often overlooked in the sweeping laudatory descriptions of the model, which extol the generality of the approach versus more highly restrictive models such as multinomial logit. In fact, users of probit must in the final analysis make modeling assumptions which are analogous to choosing among various alternative tree structures in the nested multinomial logit or tree extreme value models.

To perform this exercise, however, knowledge of the number of estimable parameters is still not enough. Arbitrarily selected specifications for \( \Sigma_e \) which contain the prescribed number of estimable parameters may not be identified, since the transformation between the parameters in the specification and the "observable" \( C_f \) parameters may not be unique and invertible. Examples are given which demonstrate that an analysis is essential for ensuring valid results. Specification errors are common in the literature and these issues appear to not be generally understood. The results presented here provide useful guidelines for those practitioners seeking to apply probit models.