Bayesian Analysis of Multivariate Normal Models when Dimensions are Absent

Robert Zeithammer

University of Chicago

Peter Lenk

University of Michigan

http://webuser.bus.umich.edu/plenk/downloads.htm

Outline

- Motivation
- Multivariate Regression
- HB Multivariate Regression
- HB Multinomial Probit Model
- Choice-Based Conjoint (CBC) Example

Motivation

- Absent dimensions occur in multivariate problems when one or more dimensions are completely unobserved for some sampling units
- It differs from usual missing data problems in that both the independent and dependent variables are unobserved
- Problem is so pervasive that researchers may not recognize that they have absent dimensions

Examples

- Not all stores carry all brands in every time period
  - Sales are missing for absent dimensions
  - Marketing mix is missing
- Not all choice sets include every brand in CBC Study
- Different schools offer different educational programs

So What?

- Imputing both independent and dependent observations for absent dimension is ill-poised problem in many contexts
- Likelihood function is well-defined, but
  - Multivariate observations have different lengths
  - Inverted Wishart is no longer conjugate for the error covariance matrix
  - Could do it with Metropolis, but that is not fun

Common Kludge # 1

- Restrict analysis to subset of dimensions that are present across all units
- Example: brand demand study
  - Exclude small-share brands
  - Focus on national brands and store brand
  - Distorts market analysis
- Example: educational outcome study
  - Focus on common set of programs
  - Potentially biases outcomes

Common Kludge # 2

- Ignore error correlations
- Example: CBC Brand Study
  - More brands in study than alternatives in choice sets
  - Distorts estimated heterogeneity
  - Misleading market share simulations
  - IIA worries

Common Kludge # 3

- Pool absent dimensions into “Other” dimension
- Keeps full covariance
- Meaning of “Other” is problematic
  - Demand for “Other”?
  - Marketing mix for “Other”?

Simple Solution

- In MCMC impute the missing *error term* for the absent dimensions
- Continue as though you have the full data set
- Adds about three lines of code
- Adds an indicator for absent dimensions to data structure

Multivariate Regression

- Model: for $i = 1, \ldots, n$

$$Y_i \;=\; X_i\beta + \epsilon_i \text{ with } \epsilon_i \sim N_m(0, \Sigma)$$

- Priors $\beta \sim N_p(b_0, V_0)$ and $\Sigma \sim IW_m(f_0, S_0)$
- $\mathcal{A}(i)$ is set of indices for the absent dimensions with $\#\mathcal{A}(i) = m_i$
- $\mathcal{P}(i)$ is set of indices for the present dimensions with $\#\mathcal{P}(i) = m - m_i$

MCMC: Initial Assignment

- Initialization of absent dimensions
  - $Y_{\mathcal{A}(i)} \leftarrow 0$
  - $X_{\mathcal{A}(i)} \leftarrow 0$
- Setting $X_{\mathcal{A}(i)}$ to zero facilitates draws of the regression coefficients from their full conditional distributions

MCMC: Absent Residuals

- Present residuals: $R_{\mathcal{P}(i)} = Y_{\mathcal{P}(i)} - X_{\mathcal{P}(i)}\beta$
- Absent residuals from conditional normal

$$R_{\mathcal{A}(i)}|R_{\mathcal{P}(i)}, \Sigma, \beta \sim N_{m-m_i}(\mu_{\mathcal{A}(i)|\mathcal{P}(i)}, \Sigma_{\mathcal{A}(i)|\mathcal{P}(i)})$$

  - Conditional mean

$$\mu_{\mathcal{A}(i)|\mathcal{P}(i)} = \Sigma_{\mathcal{A}(i),\mathcal{P}(i)}\Sigma^{-1}_{\mathcal{P}(i),\mathcal{P}(i)}R_{\mathcal{P}(i)}$$

  - Conditional covariance

$$\Sigma_{\mathcal{A}(i)|\mathcal{P}(i)} = \Sigma_{\mathcal{A}(i),\mathcal{A}(i)} - \Sigma_{\mathcal{A}(i),\mathcal{P}(i)}\Sigma^{-1}_{\mathcal{P}(i),\mathcal{P}(i)}\Sigma_{\mathcal{P}(i),\mathcal{A}(i)}$$

MCMC: Update Assignment

- $Y_{\mathcal{A}(i)} \leftarrow R_{\mathcal{A}(i)}$
- $X_{\mathcal{A}(i)} \leftarrow 0$

MCMC: β and Σ

- $\beta|$ Rest $\sim N_p\left(b_n, V_n\right)$
  - $V_n = \left(V_0^{-1} + \sum_{i=1}^n X_i' \Sigma^{-1} X_i\right)^{-1}$
  - $b_n = V_n\left(V_0^{-1} b_0 + \sum_{i=1}^n X_i \Sigma^{-1} Y_i\right)$
- $\Sigma|$Rest $\sim IW_m(f_n, S_n)$
  - $f_n = f_0 + n$
  - $S_n = S_0 + \sum_{i=1}^n \left(Y_i - X_i\beta\right)\left(Y_i - X_i\beta\right)'$
- Same code as though all dimensions are present because

Two Simulations

- m = 3; n = 500, and p = 2
- One dimension is absent for each observation
- Simulation A
  - Observe all pairs of present dimensions
  - {1,2}, {1,3}, and {2,3}
- Simulation B
  - Only observe pairs {1,2} and {2,3}
  - No sample information about $\sigma_{1,3}$

Regression Coefficients

Recovers true values

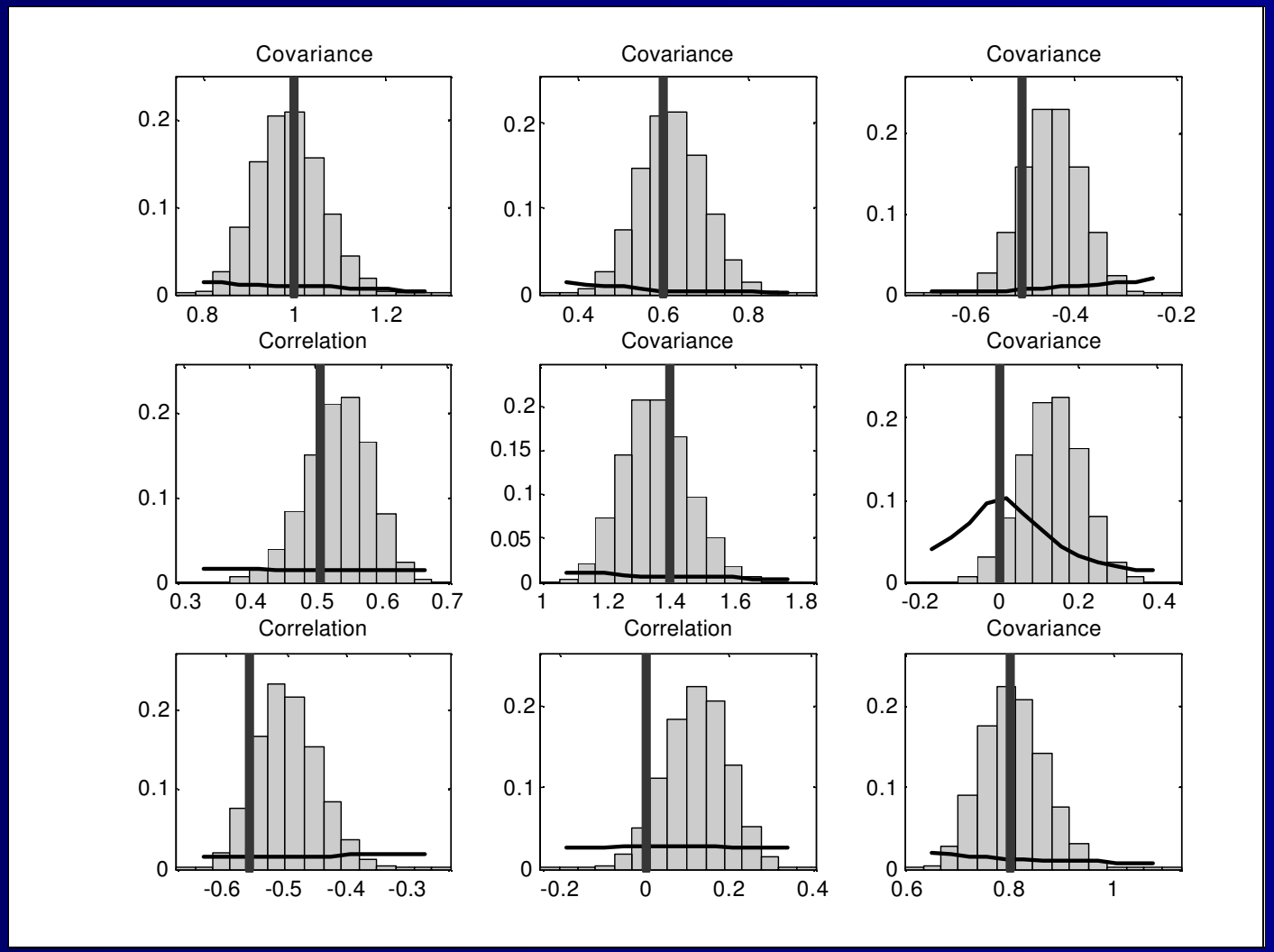
		Simulation A		Simulation B	
Coefficient	True	Mean	STD	Mean	STD
β_1	1.0	1.057	0.036	1.062	0.042
β_2	-1.0	-0.958	0.033	-0.953	0.040

Error Variance

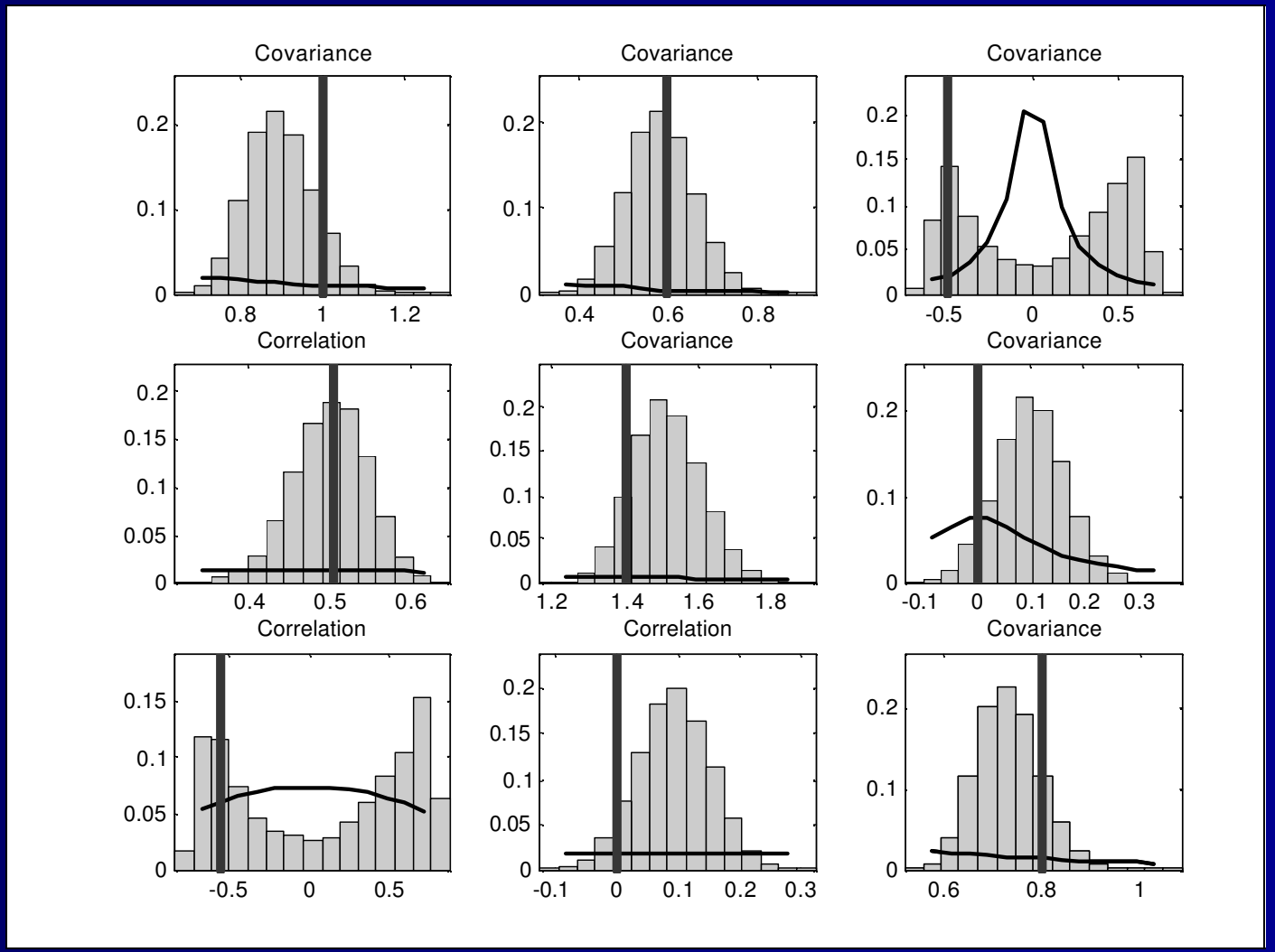
Estimate of $\sigma_{1,3}$ for Simulation B is based on prior, but other parameters are recovered

		Simulation A		Simulation B	
Covariance	True	Mean	STD	Mean	STD
$\sigma_{1,1}$	1.0	0.990	0.074	0.900	0.082
$\sigma_{1,2}$	0.6	0.622	0.078	0.586	0.076
$\sigma_{1,3}$	-0.5	-0.445	0.059	0.072	0.451
$\sigma_{2,2}$	1.4	1.358	0.105	1.517	0.096
$\sigma_{2,3}$	0.0	0.132	0.080	0.100	0.064
$\sigma_{3,3}$	0.8	0.809	0.062	0.724	0.065

Simulation A: Error Variance



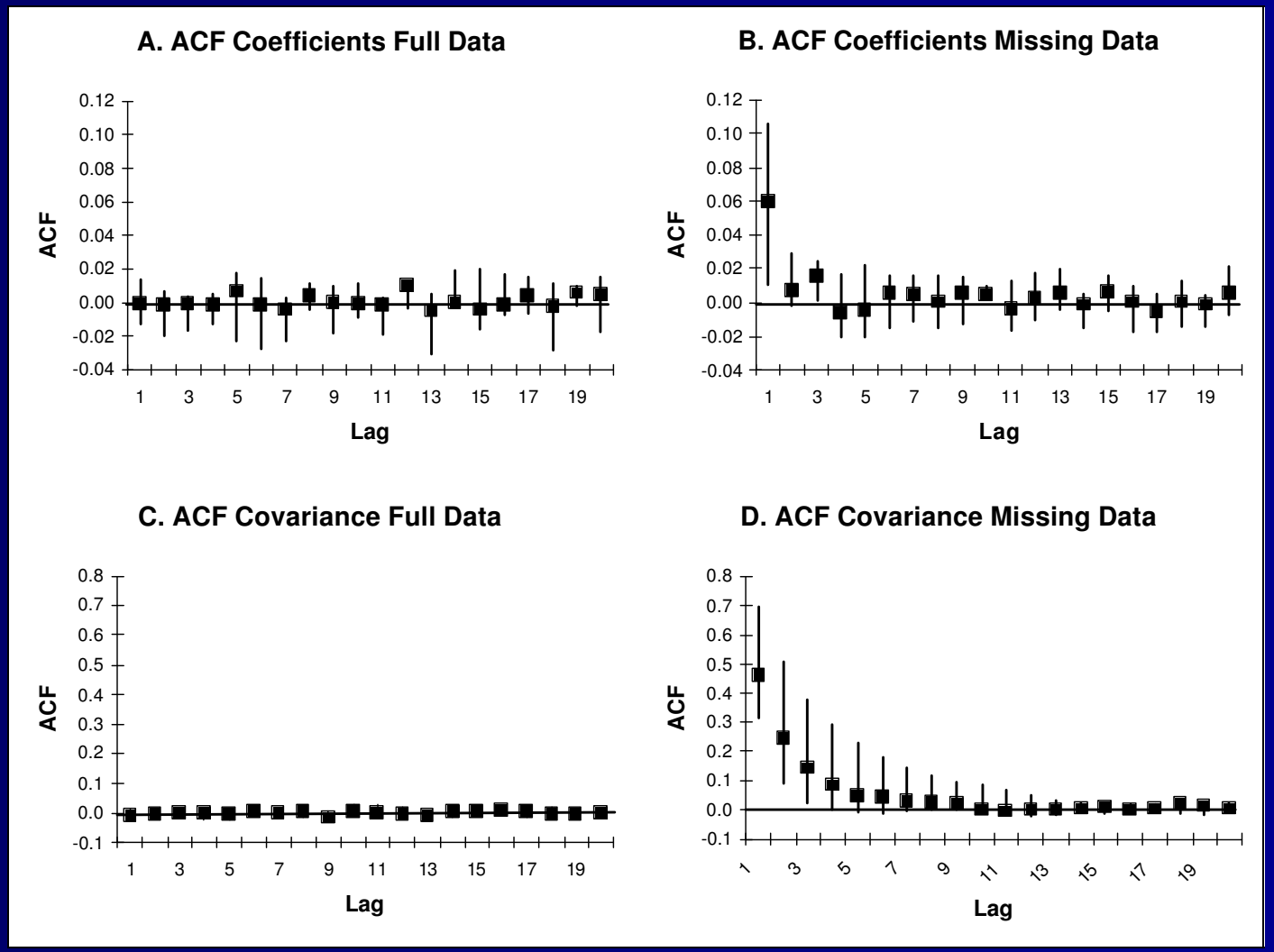
Simulation B: Error Variance



Mixing

- Pay a small price in mixing of the MCMC chain
- Simulation
  - n = 500; m = 3; p = 4
  - Full data set
  - $\frac{1}{3}$ of the dimensions were randomly deleted
  - Posterior means are close for full and absent cases
  - Posterior standard deviations are small for full case
  - ACF on next slide

Full versus Absent ACF



HB Multivariate Regression

- Model: for $j = 1, \ldots, n_i$ and $i = 1, \ldots, N$

$$\begin{aligned} Y_{ij} &= X_{ij}\beta_i + \epsilon_{ij} \text{ with } \epsilon_i \sim N_m(0, \Sigma) \\ \beta_i &= \Theta' z_i + \delta_i \text{ with } \delta_i \sim N_p\left(0, \Lambda\right) \end{aligned}$$

- Priors

$$\begin{aligned} \Sigma &\sim IW_m(f_0, S_0) \\ \Lambda &\sim IW_p(g_0, T_0) \\ \vec{\Theta}' &\sim N_{pq}\left(U_0, V_0\right) \end{aligned}$$

Analysis

- Full conditional distribution of the residuals $R_{\mathcal{A}(i,j)}$ for the absent dimensions has a conditional normal distribution given $R_{\mathcal{P}(i,j)}$
- Simulation
  - m = 4; p = 5, and q = 3 (covariate z_i)
  - N = 500 and $11 \leq n_i \leq 20$
  - One or two absent dimensions for each observation

Fit Statistics for β_i

	Correlation	RMSE
Intercept 1	0.972	1.824
Intercept 2	0.732	1.970
Intercept 3	0.692	2.140
Intercept 4	0.864	2.319
X1	0.998	0.364
X2	0.969	0.662

Error Variance

True	Y1	Y2	Y3	Y4
Y1	1.0	0.1	0.0	1.0
Y2	0.1	4.0	0.0	4.1
Y3	0.0	0.0	9.0	0.0
Y4	1.0	4.1	0.0	21.0
Bayes	**Y1**	**Y2**	**Y3**	**Y4**
Y1	1.004	0.068	0.154	0.935
Y2	0.068	4.052	0.180	4.111
Y3	0.154	0.180	9.131	0.166
Y4	0.935	4.111	0.166	21.529

Explained Heterogeneity Θ

True	CNST 1	CNST 2	CNST 3	CNST 4	X1	X2
CNST	-15.0	-5.0	5.0	20.0	-5.0	3.0
Z1	2.0	1.0	0.0	-2.0	1.0	-0.2
Z2	-1.0	-0.5	0.0	1.0	-0.2	0.5
Bayes	**CNST 1**	**CNST 2**	**CNST 3**	**CNST 4**	**X1**	**X2**
CNST	-14.778	-6.497	5.521	18.754	-4.168	-2.199
Z1	1.745	0.920	-0.203	-2.148	0.951	0.282
Z2	-0.798	-0.295	0.070	1.333	-0.186	0.530

Unexplained Heterogeneity Λ

True	CNST 1	CNST 2	CNST 3	CNST 4	X1	X2
CNST 1	0.250	-0.500	0.750	0.000	0.125	-0.150
CNST 2	-0.500	2.000	-1.000	0.000	-0.750	-0.700
CNST 3	0.750	-1.000	4.750	0.000	1.625	-0.875
CNST 4	0.000	0.000	0.000	4.000	0.000	0.000
X1	0.125	-0.750	1.625	0.000	7.563	2.975
X2	-0.150	-0.700	-0.875	0.000	2.975	11.093
Bayes	**CNST 1**	**CNST 2**	**CNST 3**	**CNST 4**	**X1**	**X2**
CNST 1	0.277	-0.002	-0.107	0.251	0.432	0.586
CNST 2	-0.002	2.160	-1.571	-0.421	0.034	0.252
CNST 3	-0.107	-1.571	3.363	-1.207	2.255	-0.377
CNST 4	0.251	-0.421	-1.207	3.951	0.726	0.678
X1	0.432	0.034	2.255	0.726	8.586	3.281
X2	0.586	0.252	-0.377	0.678	3.281	10.414

HB Multinomial Probit

- Varying choice sets $\mathcal{P}(i,j)$
- Random Utility Model

$$\begin{aligned} Y_{ij} &= X_{ij}\beta_i + \epsilon_{ij} \text{ with } \epsilon_i \sim N_{\mathcal{P}(i,j)}(0,\Sigma) \\ \beta_i &= \Theta' z_i + \delta_i \text{ with } \delta_i \sim N_p(0,\Lambda) \end{aligned}$$

- Generate $Y_{\mathcal{P}(i,j)}$ given $R_{\mathcal{A}(i,j)}$ to satisfy order condition that the utility for the observed choice exceeds the other
- Generate $R_{\mathcal{A}(i,j)}$ given $Y_{\mathcal{P}(i,j)}$: no side conditions

CBC Experiment

- Sawtooth Software Data
- 326 IT purchasing managers
- PC Profiles
  - 5 brands of PC
  - 4 Product attributes with 3 levels each
  - 4 levels for Price
- 8 Choice tasks per subject
- 3 Profiles per task plus “None”
- Firm and purchasing manager covariates

Models

- Model 1: impute absent dimensions
  - Errors associated with 5 brand concepts
  - 3 brands in each choice task
  - 2 absent dimensions
- Model 2: independent errors
  - Each brand has differen error variance
  - Zero covariances
- Model 3: errors go with presentation order
- Last profile held-out for predictive accuracy

Error Variances: Model 1

	Brand A	Brand B	Brand C	Brand D	Brand E
Brand A	0.889	0.174	-0.156	**-0.716**	0.040
Brand B	0.174	0.860	0.055	0.037	**-0.564**
Brand C	-0.156	0.055	0.961	**-0.247**	**-0.754**
Brand D	**-0.716**	0.037	**-0.247**	0.875	0.135
Brand E	0.040	**-0.564**	**-0.754**	0.135	1.000

Error Variances: Models 2 and 3

Model 2	Brand A	Brand B	Brand C	Brand D	Brand E
Brand A	1.042	0.000	0.000	0.000	0.000
Brand B	0.000	1.041	0.000	0.000	0.000
Brand C	0.000	0.000	1.053	0.000	0.000
Brand D	0.000	0.000	0.000	1.036	0.000
Brand E	0.000	0.000	0.000	0.000	1.000
Model 3	**Order 1**	**Order 2**	**Order 3**		
Order 1	1.386	-0.569	-0.617		
Order 2	-0.569	1.107	-0.535		
Order 3	-0.617	-0.535	1.000		

Estimation Results

- Estimated partworths and explained heterogeneity tend to be similar for all three models
- Pattern of “important” factors differ
- Unexplained heterogeneity is much larger for Model 2 than Models 1 and 3
  - Assuming independent errors seems to move error variation to partworth heterogeneity

Hold-Out Predictive Performance

	Hit Rate	Improvement
Model 1	56.6%	
Model 2	52.2%	8.4%
Model 3	48.8%	16.1%
	Brier Score	Reduction
Model 1	0.377	
Model 2	0.479	21.4%
Model 3	0.508	25.8%

Conclusion

- Absent dimensions occur frequently
- Complicates estimation, especially of variances
- Ad hoc approaches
  - “Data washing”
  - Assume it away with independence
- Imputing absent residuals is effective and easy