

# New Developments in Analytical Methods for Activity-Based Modeling

**Chandra R. Bhat**

Department of Civil Engineering,  
The University of Texas at Austin.

*Workshop on*

“Activity-Based Approaches: Theory, Methods, Data, and Applications”

Transportation Research Board 84<sup>th</sup> Annual Meeting

January 2005, Washington D.C.


## Overview: A Revolution in Choice Modeling

- The last decade: A very fertile period in the field of choice models
  - ✓ New GEV and “mixed” structures for discrete choice modeling
  - ✓ Flexible models for continuous (duration) choices
  - ✓ Advanced models for discrete-continuous choices
  - ✓ Simultaneous modeling of multiple choices
  - ✓ Substantial progress in simulation-estimation techniques
- Ability to specify and estimate practically *any* behavioral model structure

## Discrete Choice Models

### Mixed Generalized Extreme Value (MGEV) Model

#### Application:

Accommodating spatial correlations and response heterogeneity in residential location-choice model (The mixed-spatially correlated logit, MSCL) 

### Multidimensional Mixed Ordered Response Logit (MMORL) Model


#### Application:

Simultaneous choice of activity-episode frequency for multiple activity types 

# Continuous Choice Models

## Flexible Hazard Duration Models

### Application:

Modeling inter-shopping durations  
accommodating response heterogeneity and  
endogenous explanatory variables 

## Multivariate Hazard Duration Models

### Application:

Simultaneous modeling of inter-episode durations  
for multiple activity types 

## Discrete-Continuous Choice Models


Joint Mixed-Logit/Hazard-Duration Model

Application:

Modeling generation and allocation of household maintenance activities 

Mixed Multiple Discrete-Continuous Extreme Value (MMDCEV) Model

Application:

Choice of one or more activity types and duration of participation in each type 


# Simulation-Estimation Techniques

## Pseudo Monte-Carlo (PMC) Methods

- Computes the average of the integrand over a sequence of “random” points over the domain of integration
- Pseudo-random sequences used in implementations
- Slow asymptotic convergence
- Applicable for a wide class of integrands
- Integration error can be easily determined

# Simulation-Estimation Techniques

## Quasi Monte-Carlo (QMC) Methods

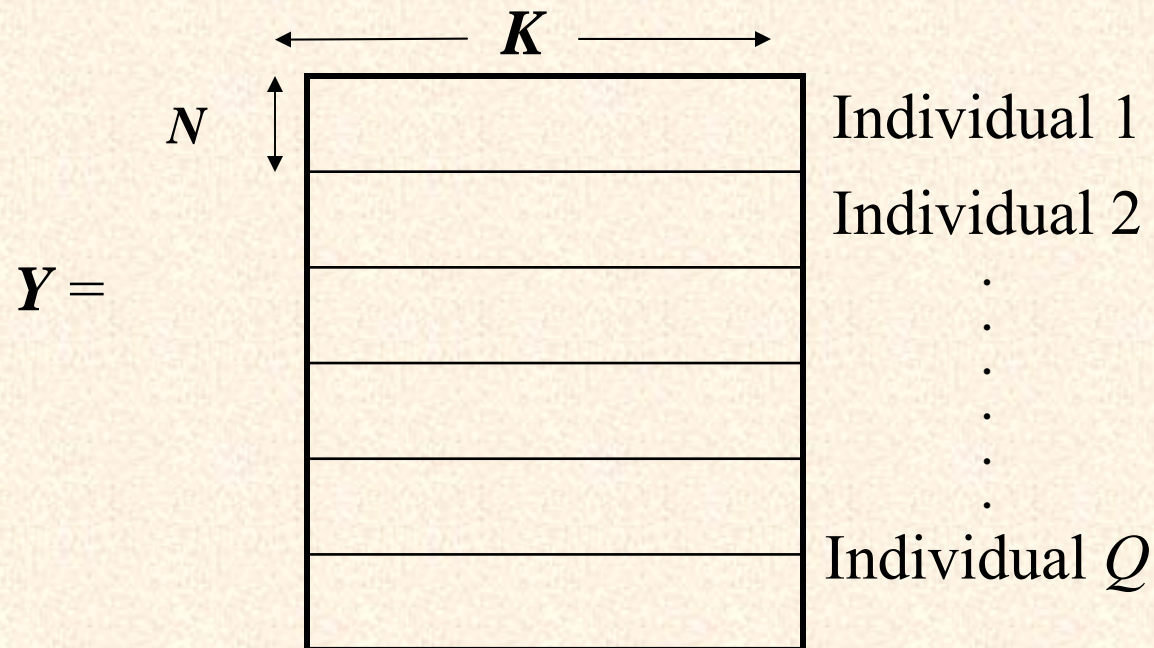
- Computes the average of the integrand over a non-random, more uniformly distributed, sequence of points over the domain of integration 
- Quasi-random sequences used in implementation
  - Halton Sequences
  - Faure Sequences
  - Latin Hypercube Sampling Sequences
- Faster convergence than PMC methods
- Substantially fewer number of draws required

# Simulation-Estimation Techniques

## Quasi Monte-Carlo (QMC) Methods

Simulation approach for choice models using QMC sequences:

- Generate a Halton matrix  $Y$  of size  $G \times K$ ,  $G = N * Q$



- Evaluate contribution of each observation by averaging across  $N$  draws

# Simulation-Estimation Techniques

## Quasi Monte-Carlo (QMC) Methods: Scrambling and Randomization

**Scrambling** breaks correlations in higher dimensions

Methods:

- ✓ Braaten-Weller scrambling for Halton sequences
- ✓ Random digit scrambling for Faure sequences
- ✓ Random linear scrambling for Faure sequences

**Randomization** enables estimation of integration error

Methods:

Tuffin's randomization 

## Summary

- The field of econometric modeling has seen a quantum jump in recent years in our ability to model
  - ✓ Discrete choices
  - ✓ Multidimensional category choices
  - ✓ Multidimensional durations
  - ✓ Discrete-continuous choices
  - ✓ Multiple discrete-continuous choices

## Summary

- A sense of absolute control over the behavioral structures that can be specified
- Value of recent contributions:  
Ability to better address the inevitable presence of unobserved factors, *even after adopting best systematic specifications*
- Renewed excitement in the field with anticipation of new developments on the horizon !!

**Thank**

**You!**

# The Mixed Spatially Correlated Logit (MSCL) Model

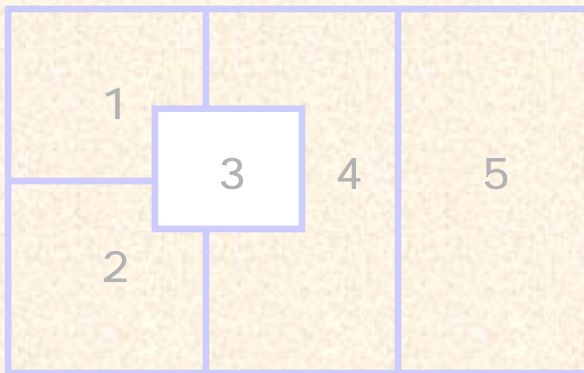
Spatial correlations accommodated via the GEV error structure

$$F(\varepsilon_{n,1}, \varepsilon_{n,2}, \dots, \varepsilon_{n,I}) = \exp \left\{ - \sum_{i=1}^{I-1} \sum_{j=i+1}^I \left[ (\alpha_{i,ij} e^{-\varepsilon_{n,i}})^{1/\mu} + (\alpha_{j,ij} e^{-\varepsilon_{n,j}})^{1/\mu} \right]^\mu \right\}$$

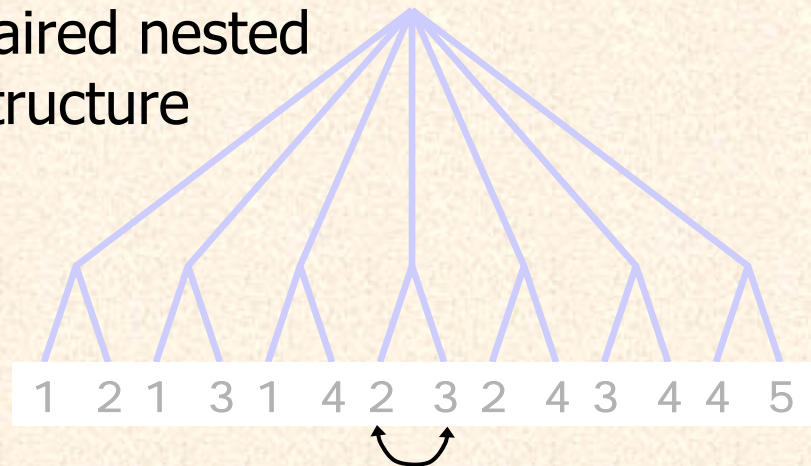
- ✓ Every pair of adjacent zones is a “paired nest”
- ✓ Greater the number of zones adjacent to a given zone (i), lesser is the correlation between i and any of the adjacent zones (j) [captured by  $\alpha_{i,ij}$ ]
- ✓ The correlation between any pair of zones decreases with the increase in the dissimilarity parameter [ $\mu$ ]
- ✓ Closed-form analytic expression for the probability function (if there are no random parameters)

# The Mixed Spatially Correlated Logit (MSCL) Model

## Example



Paired nested structure



For zone  $i = 4$

$$\mu_{ij} = \mu, \forall (i, j)$$

$$\alpha_{i,j} = \frac{\omega_{ij}}{\sum_k \omega_{ik}}$$

$$\sum_j \alpha_{i,j} = 1$$

$j$	$\omega_{ij}$	$\alpha_{i,j}$
1	1	0.25
2	1	0.25
3	1	0.25
5	1	0.25

# The Mixed Spatially Correlated Logit (MSCL) Model

Response heterogeneity captured via a mixing distribution

Conditional probability function:

$$P_{n,i} | \beta = \frac{\sum_{j \neq i} (\alpha_{ij} e^{\beta' x_{n,i}})^{1/\mu} \left[ (\alpha_{i,ij} e^{\beta' x_{n,i}})^{1/\mu} + (\alpha_{j,ij} e^{\beta' x_{n,j}})^{1/\mu} \right]^{\mu-1}}{\sum_{k=1}^{I-1} \sum_{l=k+1}^I \left[ (\alpha_{k,kl} e^{\beta' x_{n,k}})^{1/\mu} + (\alpha_{l,kl} e^{\beta' x_{n,l}})^{1/\mu} \right]^{\mu}}$$

Unconditional probability function:

$$P_{n,i} = \int_{-\infty}^{\infty} (P_{n,i} | \beta) f(\beta | \theta) d\beta$$

Estimated using simulation-estimation techniques: the dimensionality of the integral equals the number of random  $\beta$  parameters

# The Mixed Spatially Correlated Logit (MSCL) Model

## Empirical Application

- ✓ Study area covers 98 counties in Dallas county, Texas
- ✓ Used data from the 1996 DFW household activity survey
- ✓ Sample comprises 236 households with a single worker
- ✓ Six groups of variables examined:
  - Size measures
  - Commute-related variables
  - School quality measures
  - Socio-economic & demographic variables
  - Land use variables
  - Regional accessibility variables



# The Mixed Spatially Correlated Logit (MSCL) Model

## Results

	Multinomial Logit Model		Mixed Spatially Correlated Logit Model	
	Parameter	t-statistic	Parameter	t-statistic
Logarithm of zonal area (in mile <sup>2</sup> )	0.250	2.776	0.286	3.256
Population density (in 10 persons/mile <sup>2</sup> )				
Mean	7.685	4.223	6.987	4.049
Standard Deviation	0.000	—	9.358	1.600
Percentage of zonal area occupied by multifamily housing				
Mean	-1.319	-2.063	-3.741	-2.919
Standard Deviation	0.000	—	4.541	1.914
Absolute difference between zonal median income and household income (in \$100,000)	-1.270	-2.305	-1.056	-1.762
Commute time (in 100's of minutes)				
Mean	-3.673	-2.200	-4.409	-2.441
Standard Deviation	0.000	—	6.504	1.180
Percentage zonal Hispanic population interacted with Hispanic dummy variable	1.235	1.214	1.094	1.127
Work accessibility interacted with African-American household head dummy variable	-2.921	-3.891	-2.329	-3.310
Shopping accessibility	5.809	8.350	5.098	5.759
Dissimilarity parameter	1.000	—	0.358	3.541
Number of observations	236		236	
Log-likelihood at convergence	-1013.43		-1000.93	

# The Multidimensional Mixed Ordered Response Logit (MMORL) Model

Eqn. for  
activity type 1

$$f_q^* = \alpha'x_q + \varepsilon_q + u_q,$$

$$f_q = l \quad \text{if } \delta_{l-1} < f_q^* < \delta_l$$

Eqn. for  
activity type 3

$$g_q^* = \beta'y_q + \eta_q + v_q,$$

$$g_q = m \quad \text{if } \theta_{m-1} < g_q^* < \theta_m$$

$$h_q^* = \gamma'z_q + \xi_q + w_q,$$

$$h_q = n \quad \text{if } \psi_{n-1} < h_q^* < \psi_n$$

Stop-making  
propensities

Multivariate  
normal distributed  
error terms with  
covariance  $\Sigma$

Independent  
logistically  
distributed  
error terms



# The Multidimensional Mixed Ordered Response Logit (MMORL) Model

Conditional likelihood function:

$$\begin{aligned} L_{fq} | \varepsilon_q, \eta_q, \xi_q &= \prod_l \left[ \left\{ \Lambda(\delta_l - [\alpha'x_q + \varepsilon_q]) - \Lambda(\delta_{l-1} - [\alpha'x_q + \varepsilon_q]) \right\}^{F_{ql}} \right] \\ &\quad * \prod_m \left[ \left\{ \Lambda(\theta_m - [\beta'y_q + \eta_q]) - \Lambda(\theta_{m-1} - [\beta'y_q + \eta_q]) \right\}^{G_{qm}} \right] \\ &\quad * \prod_n \left[ \left\{ \Lambda(\psi_n - [\gamma'z_q + \xi_q]) - \Lambda(\psi_{n-1} - [\gamma'z_q + \xi_q]) \right\}^{H_{qn}} \right] \end{aligned}$$

Unconditional likelihood function:

$$L_q = \int_{\varepsilon_q} \int_{\eta_q} \int_{\xi_q} (L_q | \varepsilon_q, \eta_q, \xi_q) \phi_3(\varepsilon_q, \eta_q, \xi_q) d\varepsilon_q d\eta_q d\xi_q$$



## Flexible Hazard Duration Models

$$m_q^* = \theta' h_q + \zeta_q + \nu_q$$

$$m_q = 1 \text{ if } m_q^* > 0, m_q = 0 \text{ if } m_q^* \leq 0$$

$$p_q^* = \mu' r_q + \xi_q + \omega_q$$

$$p_q = 1 \text{ if } p_q^* > 0, p_q = 0 \text{ if } p_q^* \leq 0$$

$$s_q^* = \delta' w_q + \beta' x_q \pm \zeta_q \pm \xi_q - \varpi_q + \varepsilon_q$$

$$s_q = k \text{ if } \psi_{k-1} < s_q^* < \psi_k$$

$$x_q = [m_q, p_q]$$

Use of  
mobile  
phones

Use of  
computers

Explanatory  
variables in  
inter-shopping  
duration model

Inter-shopping  
duration

# The Multivariate Hazard Duration Model

- ✓ Flexible (non-parametric baseline) structure accounting for the dynamics of activity participation across multiple days for each purpose
- ✓ Recognize the presence of unobserved individual specific attributes affecting interactivity durations
- ✓ Incorporate intra-individual variations in interactivity durations due to unobserved characteristics
- ✓ Recognize dependence among interactivity durations of each type due to unobserved individual specific factors

# The Multivariate Hazard Duration Model

Hazard function

(for person  $q$ , activity type  $m$ , and spell  $i$ )

$$\lambda_{qmi}(\tau) = \underbrace{\lambda_{m0}(\tau)}_{\text{Baseline hazard (non-parametric)}} \exp(-\beta'_m x_{qmi} - \nu_{qm} + \varpi_{qmi})$$

Baseline hazard (non-parametric)

Exogenous variables

Error term capturing unobserved individual-specific effects

$$\nu_q \sim N(0, \Omega)$$

Error term capturing unobserved intra-individual variations

$$c_{qmi} = \exp(\varpi_{qmi}) \\ \sim \text{Gamma}(1, \sigma_m^2)$$

# The Multivariate Hazard Duration Model

Conditional likelihood function  
(person  $q$ , activity type  $m$  & spell  $i$ ):

$$L_{qmi} | v_{qm} = G_{(t_{qmi}-1)} - G_{t_{qmi}}$$
$$G_{t_{qmi}} = \left[ 1 + \sigma_m^2 B_{t_{qmi}} \right]^{-\sigma_m^{-2}}$$
$$B_{t_{qmi}} = \exp \left\{ \psi_{t_{qmi}} - [\beta_m x_{qmi} + v_{qm}] \right\}$$

Unconditional likelihood function for person  $q$

$$L_q = \int_{v_q} \prod_{m=1}^{I_{qm}} \prod_{i=1} (L_{qmi} | v_{qm}) dF(v_q)$$

# The Multivariate Hazard Duration Model

- ✓ Estimated using simulation-estimation techniques: the dimensionality of the integral equals the number activity types
- ✓ Univariate draws converted to desired multivariate distribution using Cholesky decomposition of the covariance matrix
- ✓ The covariance matrix is parameterized in terms of its Cholesky decomposition in the likelihood function to ensure positive definiteness

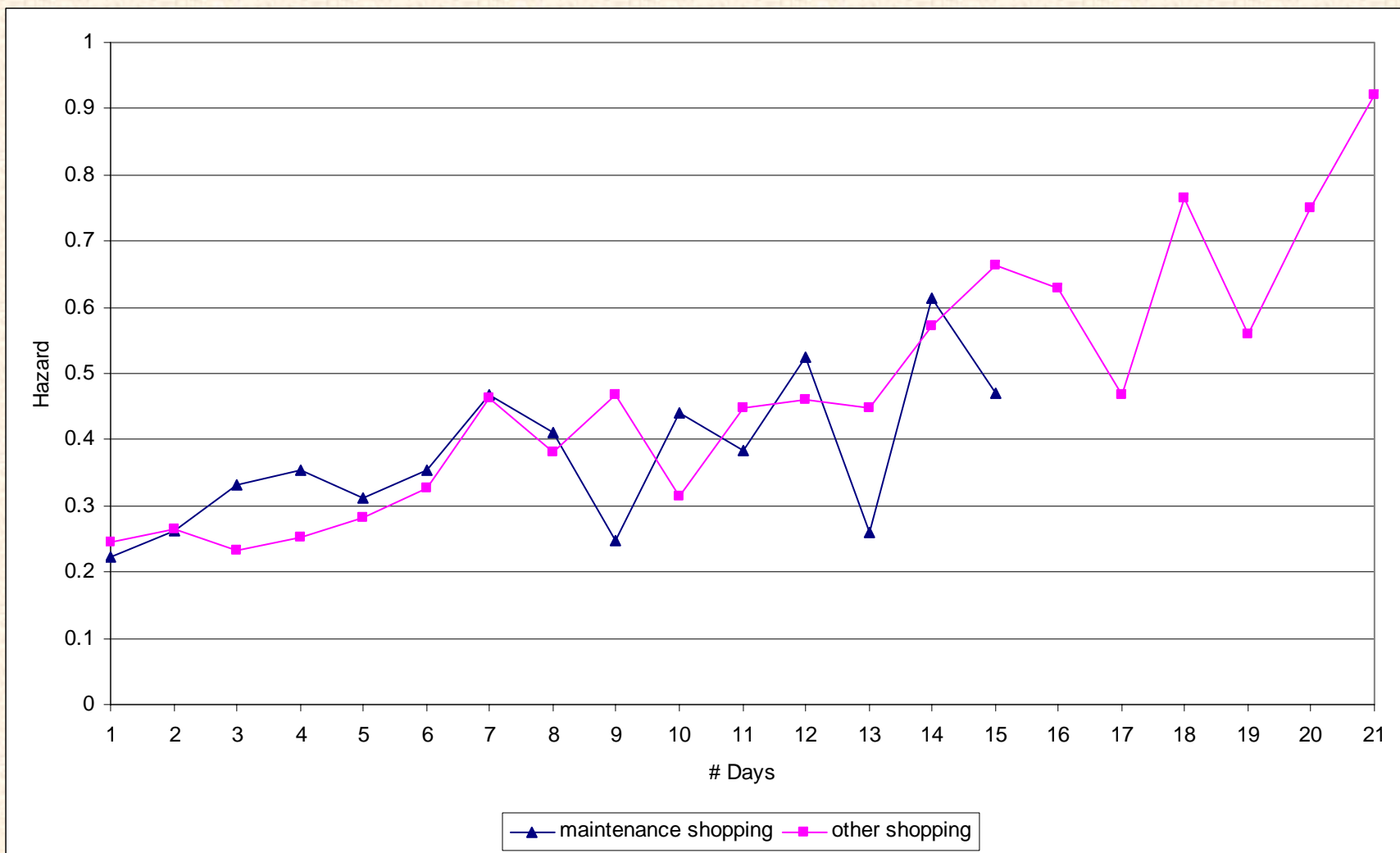
# The Multivariate Hazard Duration Model

## Empirical Application

- ✓ Six week travel survey conducted as a part of *MobiDrive* study
- ✓ Study cities: Karlsruhe (West Germany) and Halle (East Germany)
- ✓ Final sample: 361 individuals from 162 households
- ✓ Examined inter-activity durations for five activity types

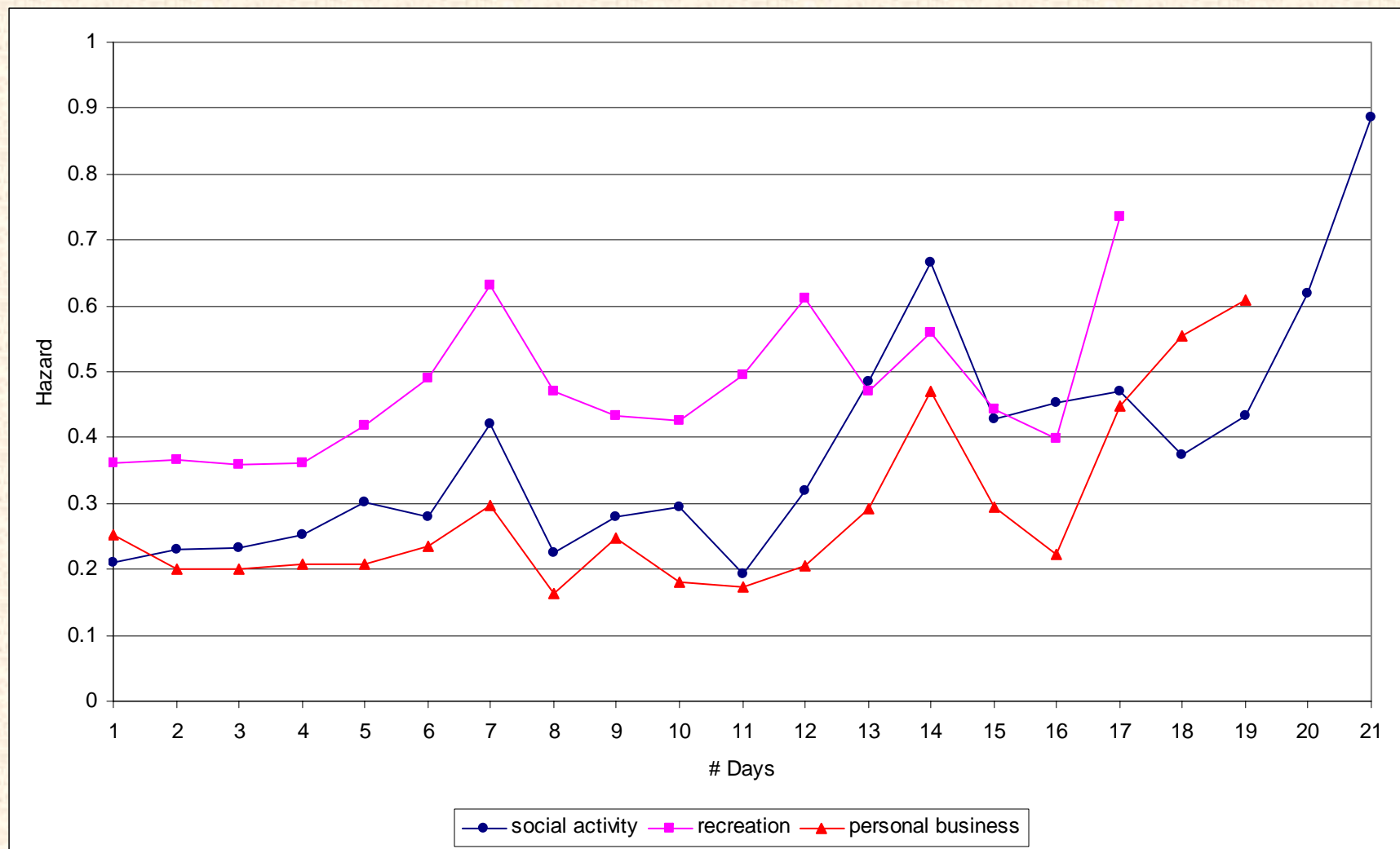
# The Multivariate Hazard Duration Model

## Estimated baseline hazard



# The Multivariate Hazard Duration Model

## Estimated baseline hazard



# The Multivariate Hazard Duration Model

## Covariance matrix of interactivity hazards

	Maintenance shopping	Non-maintenance shopping	Social activities	Recreation	Personal business
Maintenance shopping	<b>0.5621</b> <b>(5.36)</b>	<b>0.1211</b> <b>(2.61)</b>	-0.0216 <b>(-0.51)</b>	-0.0575 <b>(-1.40)</b>	<b>0.1632</b> <b>(4.06)</b>
Other shopping	-	0.064 <b>(1.54)</b>	<b>0.0433</b> <b>(1.65)</b>	0.0032 <b>(0.23)</b>	<b>0.1033</b> <b>(2.82)</b>
Social activities	-	-	<b>0.4421</b> <b>(4.13)</b>	<b>0.0738</b> <b>(2.13)</b>	<b>0.0561</b> <b>(1.66)</b>
Recreation	-	-	-	<b>0.661</b> <b>(5.68)</b>	-0.0319 <b>(-.90)</b>
Personal business	-	-	-	-	<b>0.2105</b> <b>(4.24)</b>

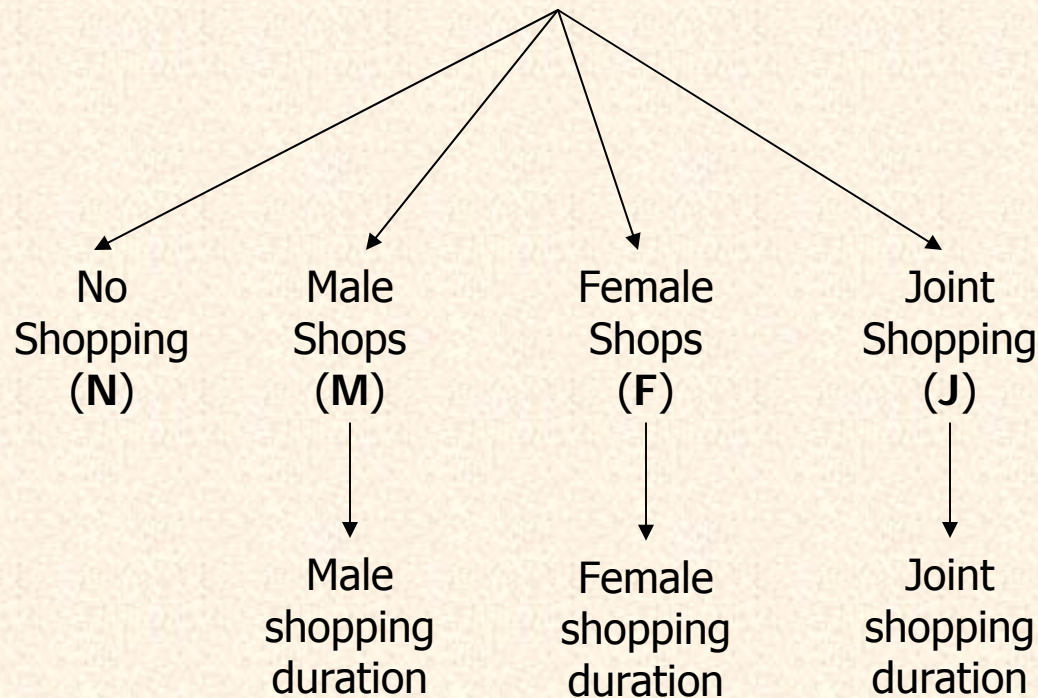


# The Multivariate Hazard Duration Model

Percentage of heterogeneity explained by  
observed and unobserved factors

Heterogeneity source	Maintenance shopping	Non-maintenance shopping	Social activities	Recreation	Personal business
Observed heterogeneity	<b>24</b>	<b>16</b>	<b>9</b>	<b>22</b>	<b>14</b>
Unobserved heterogeneity	<b>76</b>	<b>84</b>	<b>91</b>	<b>78</b>	<b>86</b>
<i>Inter-individual</i>	72	10	54	77	65
<i>Intra-individual</i>	28	90	46	23	35

# Joint Mixed-Logit Hazard-Duration Model



Household's decision to shop and task allocation: Discrete component

Shopping durations for the person(s) allocated the task: Continuous component

# Joint Mixed-Logit Hazard-Duration Model

Utility for the discrete  
choice alternatives  
( $i=N,M,F,&J$ )

$$U_{iq} = \beta_i Z_{iq} + \omega_{iq} + \varepsilon_{iq}$$

Integrated hazard  
expression for the  
shopping durations  
( $i=M,F,&J$ )

$$s_{iq}^* = \ln \int_0^{s_{iq}} \lambda_{0i}(T) dT = \gamma_i X_{iq} + \eta_{iq}$$

$$\omega_q = [\omega_{Nq}, \omega_{Mq}, \omega_{Fq}, \omega_{Jq}] \sim MVN(0, \Sigma)$$

$\varepsilon_{iq} \sim$  i.i.d. Gumbel – distributed across choice alternatives

$\eta_{iq} \sim$  extreme - value distributed  $\forall i = M, F,$  and  $J$

$\rho_i =$  correlation between  $\varepsilon_{iq}$  and  $\eta_{iq} \forall i = M, F,$  and  $J$



## Joint Mixed-Logit Hazard-Duration Model

Conditional probability of choosing not to shop (i=N)

$$P_q(R_{Nq} = 1 | \omega_q) = \frac{\exp(\beta_N Z_{Nq} + \omega_{Nq})}{\sum_{l=N,M,F,J} \exp(\beta_l Z_{lq} + \omega_{lq})}$$

Conditional probability of choosing task allocation i (i=M,F, or J) and a corresponding duration

$$P_q(R_{iq} = 1 \& M_{k_i q} = 1 | \omega_q) = \left\{ \begin{array}{l} \Phi_2 \left\{ \Phi^{-1}(F_i(\beta_i Z_{iq} + \omega_{iq} | \omega_q)), \Phi^{-1}(G(\delta_{i,k_i} - \gamma_i X_{iq}), \rho_i) \right\} - \\ \Phi_2 \left\{ \Phi^{-1}(F_i(\beta_i Z_{iq} + \omega_{iq} | \omega_q)), \Phi^{-1}(G(\delta_{i,k_i-1} - \gamma_i X_{iq}), \rho_i) \right\} \end{array} \right\}$$

## Mixed Multiple Discrete Continuous Extreme Value (MMDCEV) Model

- Accommodates multiple discreteness
- Allows diminishing marginal returns (i.e., satiation)
- Results in a simple and elegant closed form model
- Incorporates heteroskedasticity and/or correlation in unobserved characteristics

## Mixed Multiple Discrete Continuous Extreme Value (MMDCEV) Model

$$\tilde{U} = \sum_j [\exp(\beta'x_j + \varepsilon_j)] \cdot (t_j + \gamma_j)^{\alpha_j}$$

subject to time budget constraint,  $\sum_{j=1}^K t_j = T$

$\varepsilon_j$  -> idiosyncratic (unobserved) characteristics that impact the baseline utility for purpose  $j$

$\alpha_j$  -> the rate of diminishing marginal utility of investing time in activity purpose  $j$  ( $0 < \alpha_j \leq 1$ )

$\gamma_j$  -> represents translation. If this is zero for all  $j$ , only interior solution possible

## Mixed Multiple Discrete Continuous Extreme Value (MMDCEV) Model

$$\varepsilon_j = \zeta_j + \eta_j + \mu_j$$

$\zeta_j$  -> iid Gumbel distributed across alternatives

$\eta_j$  -> independently & normally across alternatives with variance  $\sigma^2_{\eta_j}$  --- introduces heteroskedasticity

$\mu = [\mu_1, \mu_2, \dots, \mu_J]$  -> Multivariate normal distributed with covariance  $\Sigma$  --- introduces error correlations across activity types

# Mixed Multiple Discrete Continuous Extreme Value (MMDCEV) Model

## Conditional likelihood function

$P(t_i^* > 0 \text{ and } t_s^* = 0; i = 2, 3, \dots, M \text{ and } s = M + 1, \dots, K)$

$$= \left[ \prod_{i=1}^M c_i \right] \left[ \sum_{i=1}^M \frac{1}{c_i} \right] \left[ \frac{\prod_{i=1}^M e^{V_i}}{\left( \sum_{j=1}^K e^{V_j} \right)^M} \right] (M-1)! \quad \text{where } c_i = \left( \frac{1 - \alpha_i}{t_i^* + \gamma_i} \right)$$

## Unconditional likelihood function

$$= \int_{\eta} \int_{\mu} \left[ \prod_{i=1}^M c_i \right] \left[ \sum_{i=1}^M \frac{1}{c_i} \right] \left[ \frac{\prod_{i=1}^M e^{V_i + \eta'w_i + \mu'z_i}}{\left( \sum_{j=1}^K e^{V_j + \eta'w_j + \mu'z_j} \right)^M} \right] (M-1)! dF(\mu | \sigma) dF(\eta | \omega)$$

# Mixed Multiple Discrete Continuous Extreme Value (MMDCEV) Model

## Empirical Application

- Used Weekend data from the BATS 2000
- Final sample: 1917 individuals 16 years or older
- Five activity purpose categories studied
  - In-home social (IHS)
  - In-home recreational (IHR)
  - Out-of-home social (OHS)
  - Out-of-home recreational (OHR)
  - Out-of-home shopping (OHSh)

# Mixed Multiple Discrete Continuous Extreme Value (MMDCEV) Model

## Satiation Parameters

Activity Type	Parameter	t-statistic
In-home social (IHS)	0.8794	3.09
In-home recreational (IHR)	0.9556	3.47
Out-of-home social (OHS)	0.7660	6.34
Out-of-home recreational (OHR)	0.7822	6.39
Out-of-home shopping (OHSh)		
Women	0.4586	7.60
Men	0.4028	7.50

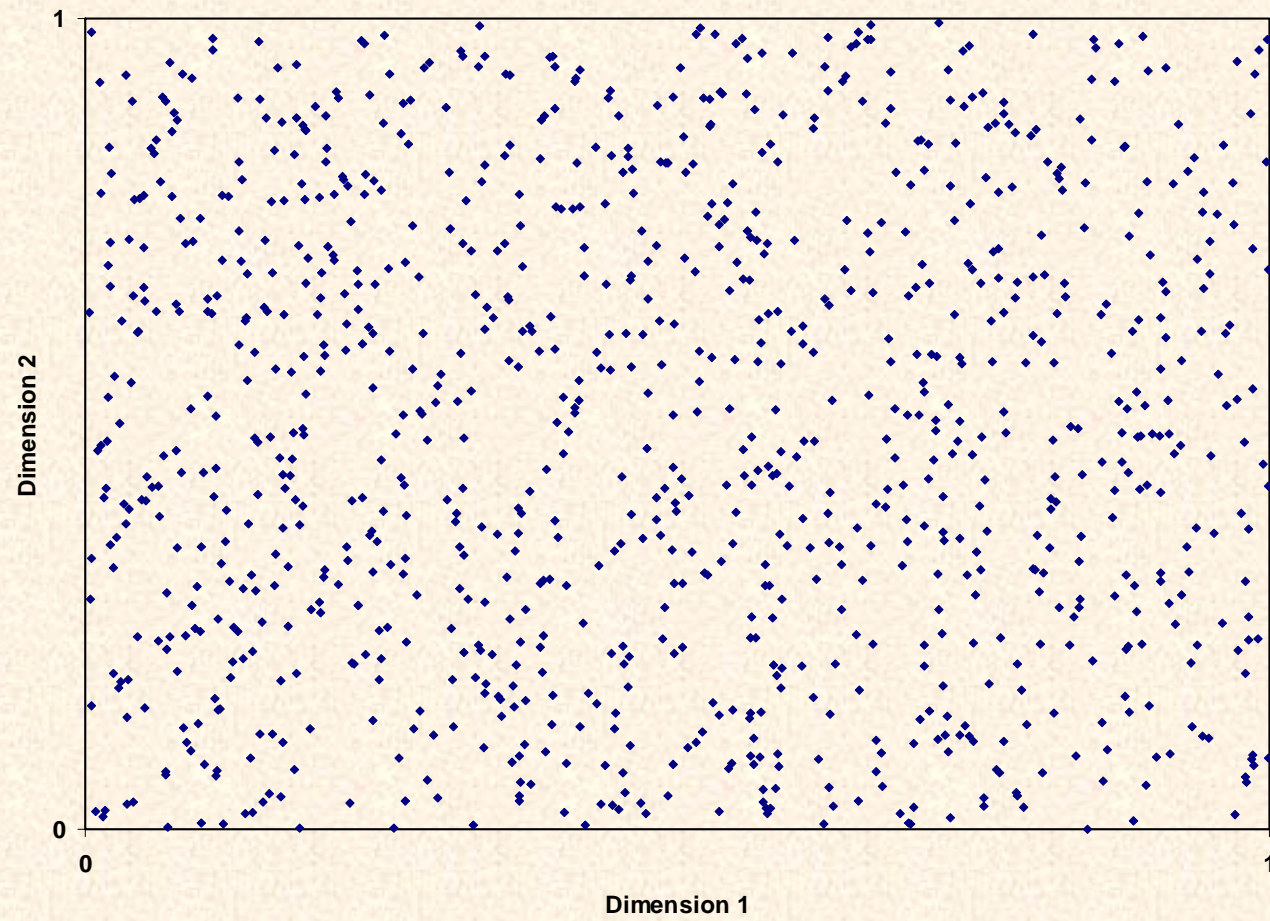


# Mixed Multiple Discrete Continuous Extreme Value (MMDCEV) Model

## Variance-Covariance matrix

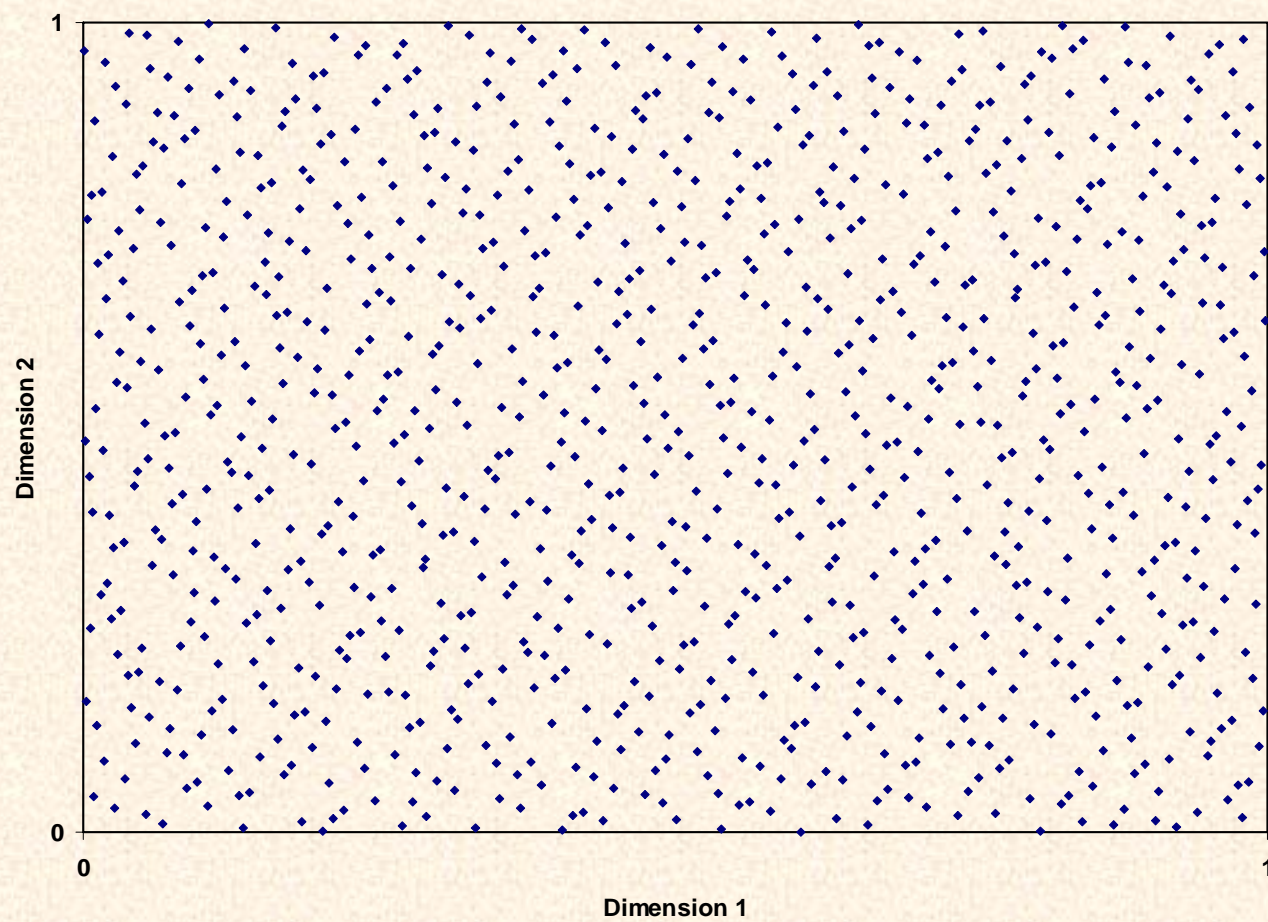
Activity Type	Activity Type				
	In-home social	In-home recreational	Out-of-home social	Out-of-home recreational	Out-of-home shopping
In-home social (IHS)	7.87 (2.98)	3.04 (2.50)	0	0	0
In-home recreational (IHR)		4.74 (3.85)	0	0	0
Out-of-home social (OHS)			11.88 (4.26)	0	0.64 (1.21)
Out-of-home recreational (OHR)				11.24 (4.11)	0
Out-of-home shopping (OHSh)					11.88 (4.26)

## 1000 Pseudo Monte Carlo Draws

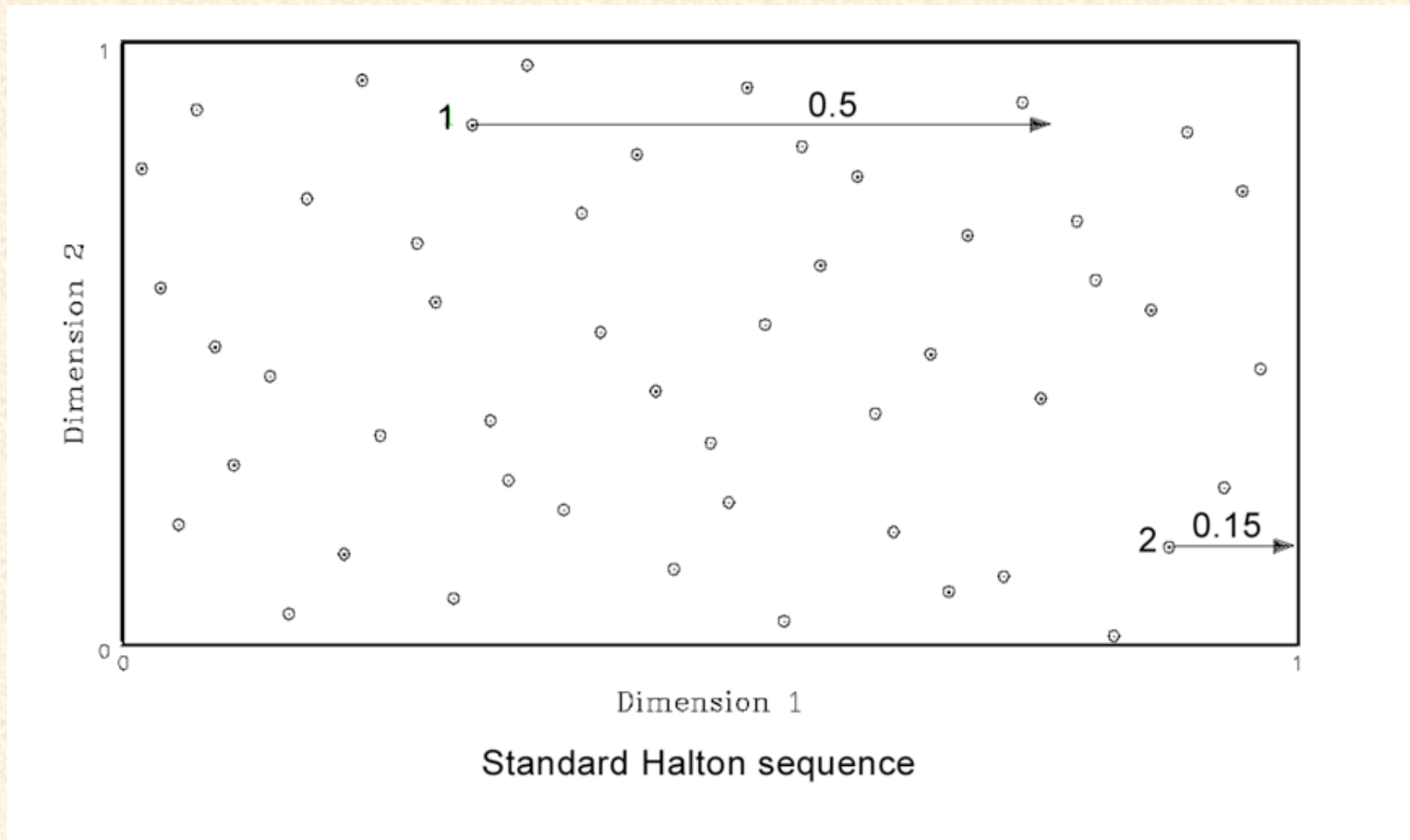




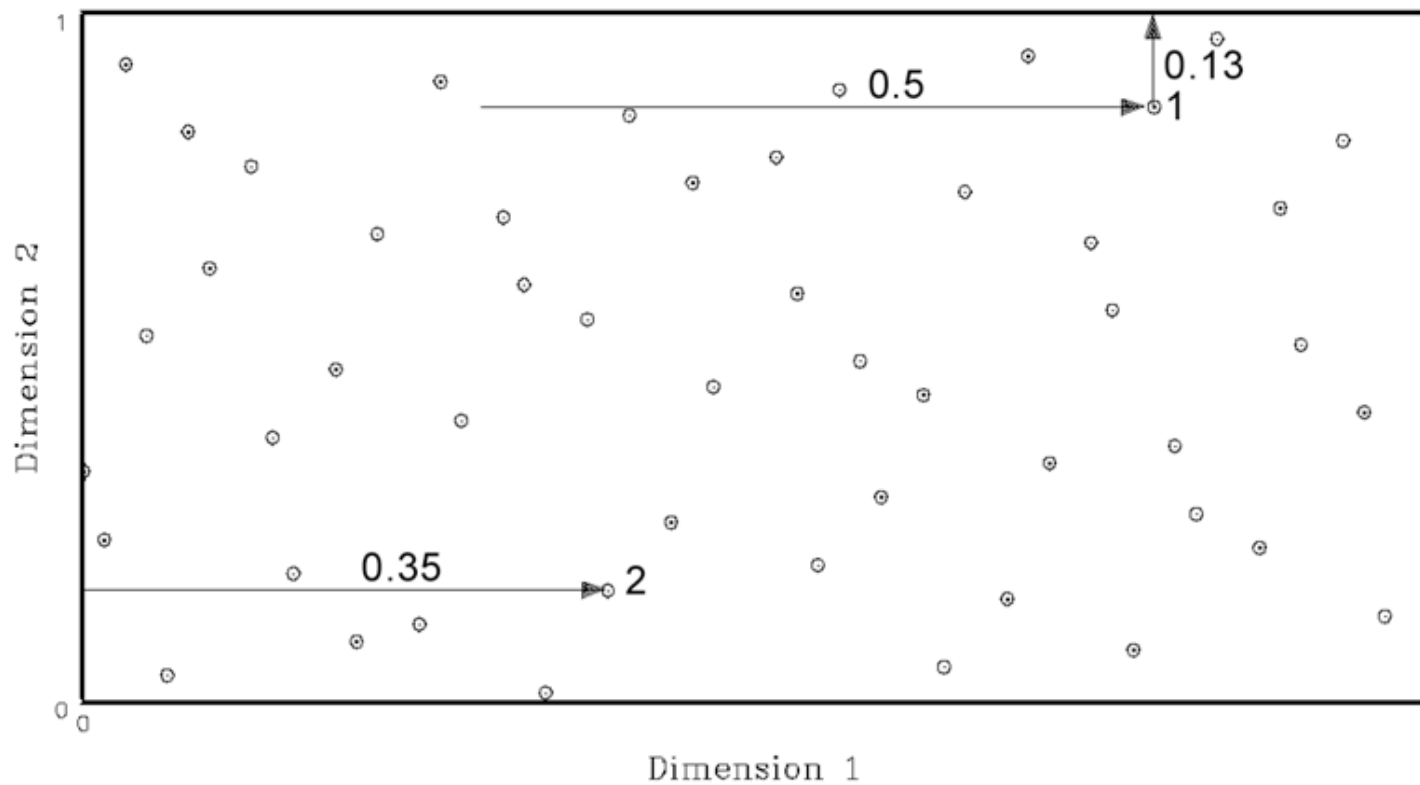
## 1000 Quasi Monte Carlo Draws



## Randomizing QMC Sequences



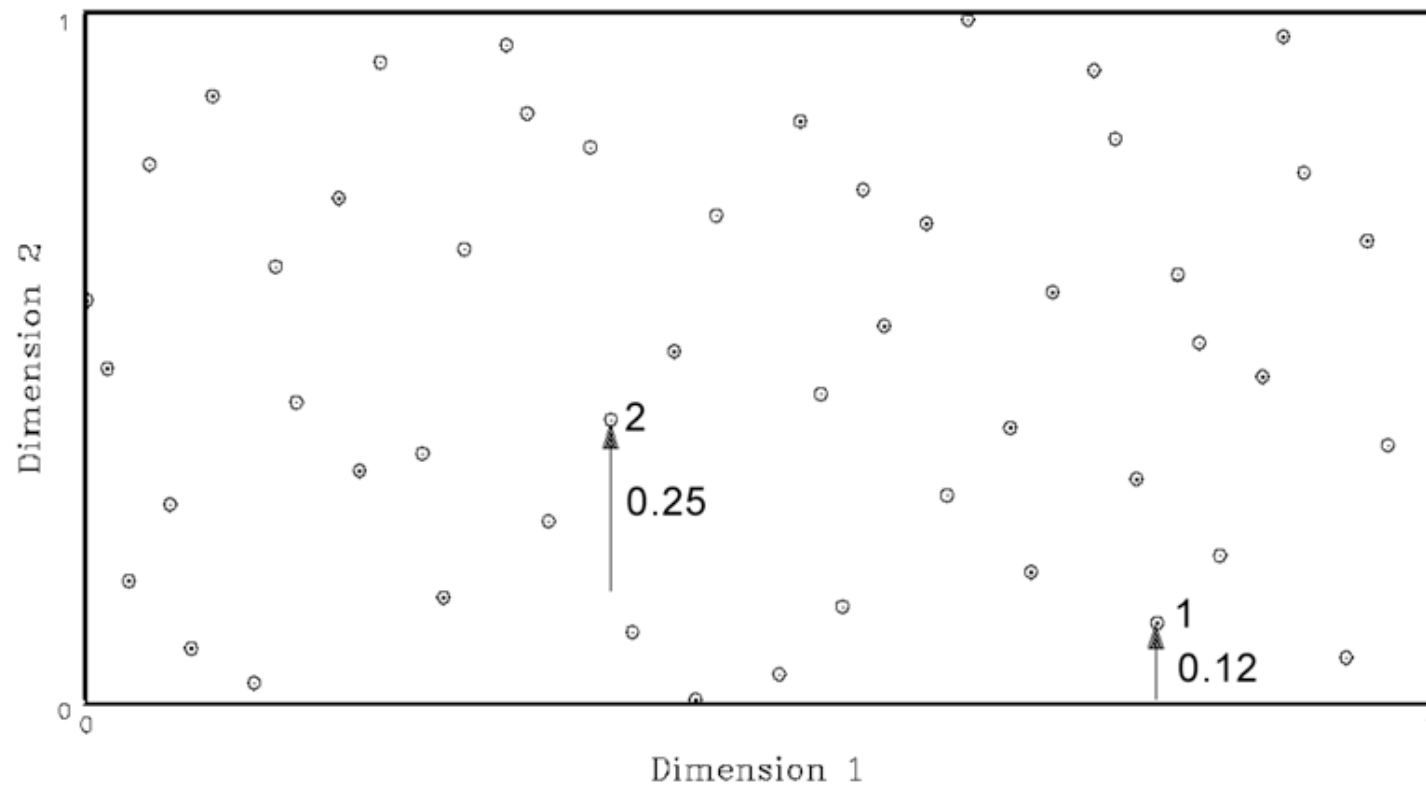
## Randomizing QMC Sequences



Standard Halton sequence shifted by 0.5 in Dimension 1



## Randomizing QMC Sequences



Standard Halton sequence shifted by 0.5 in Dimension 1 and 0.25 in Dimension 2