Continuous Time Analysis of Panel Data: An Illustration of the Exact Discrete Model (EDM)

Aaron Boulton & Pascal Deboeck

aboulton@ku.edu

University of Kansas

Goals:

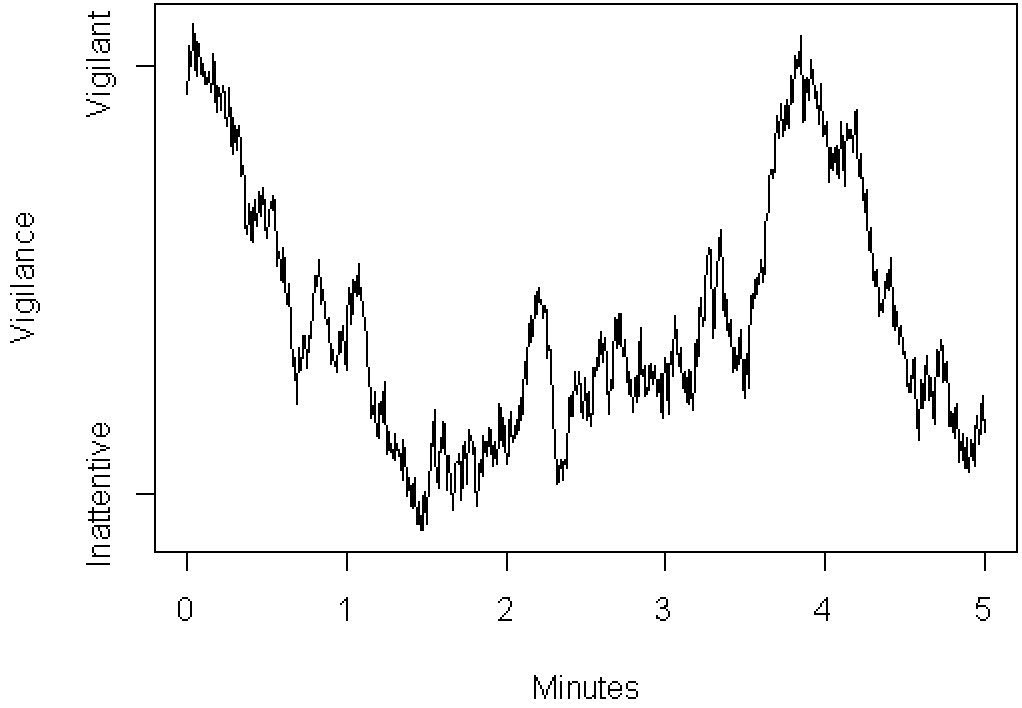
1. Discuss the problems of *discrete-time* models applied to panel data

2. Introduce the Exact Discrete Model (EDM), a *continuous-time* model based on discrete observations

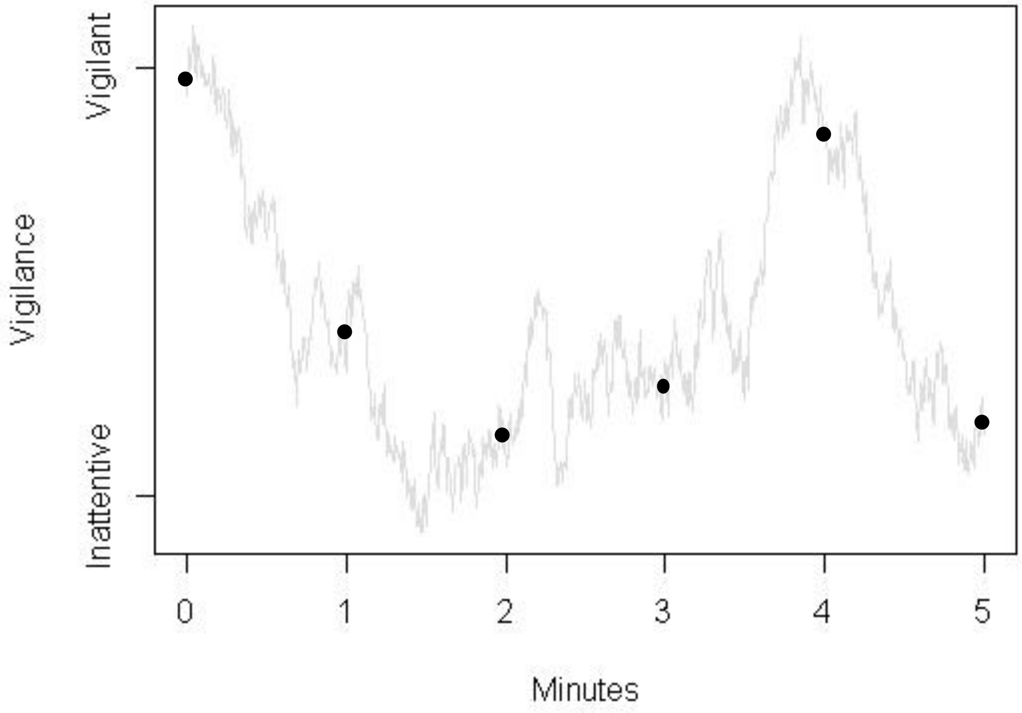
Goals:

3. Provide an illustrative example of the EDM using real data. We will estimate the model using the R program CT_SEM.R (Voelkle, Oud, Davidov, & Schmit, 2012) with data from a multi-wave panel study of the Youth Matters (YM) anti-bullying program

In reality, processes evolve continuously over time



However, researchers often are only able to observe the process at a few discrete occasions



One popular model applied to such data is the cross-lagged panel model (CLPM)

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{w}_t$$

$\mathbf{x}_t$ = vector of outcome variables

$\mathbf{A}$ = matrix of auto-regressive/cross-lagged model parameters

$\mathbf{w}_t$ = vector of prediction errors

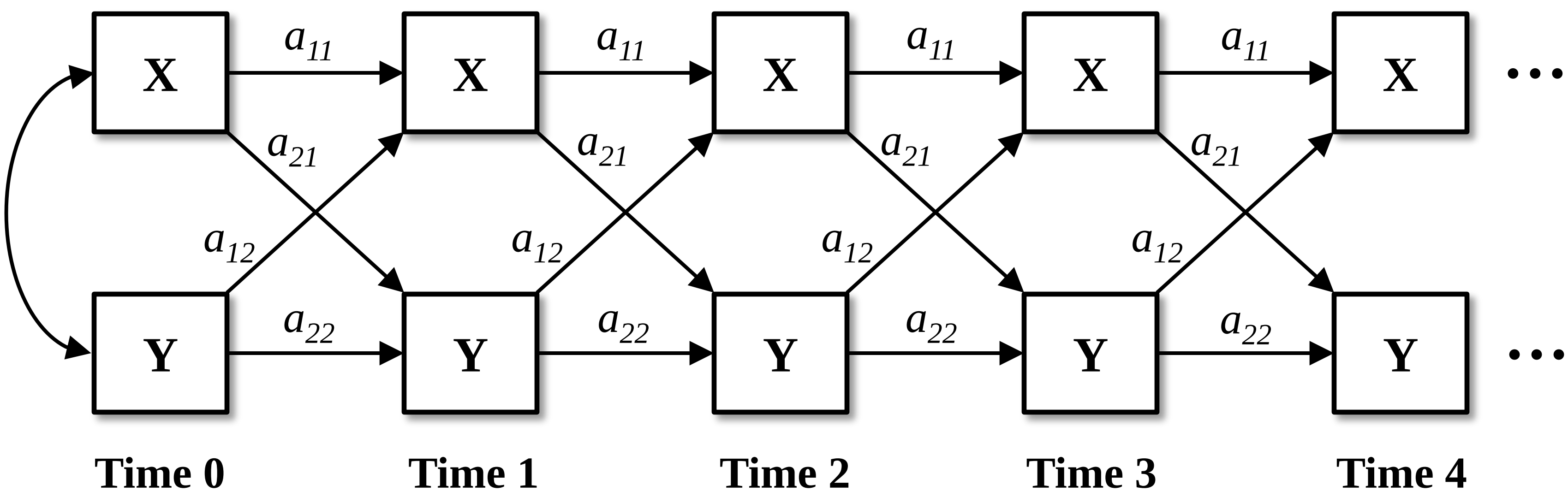
One popular model applied to such data is the cross-lagged panel model (CLPM)

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{w}_t$$

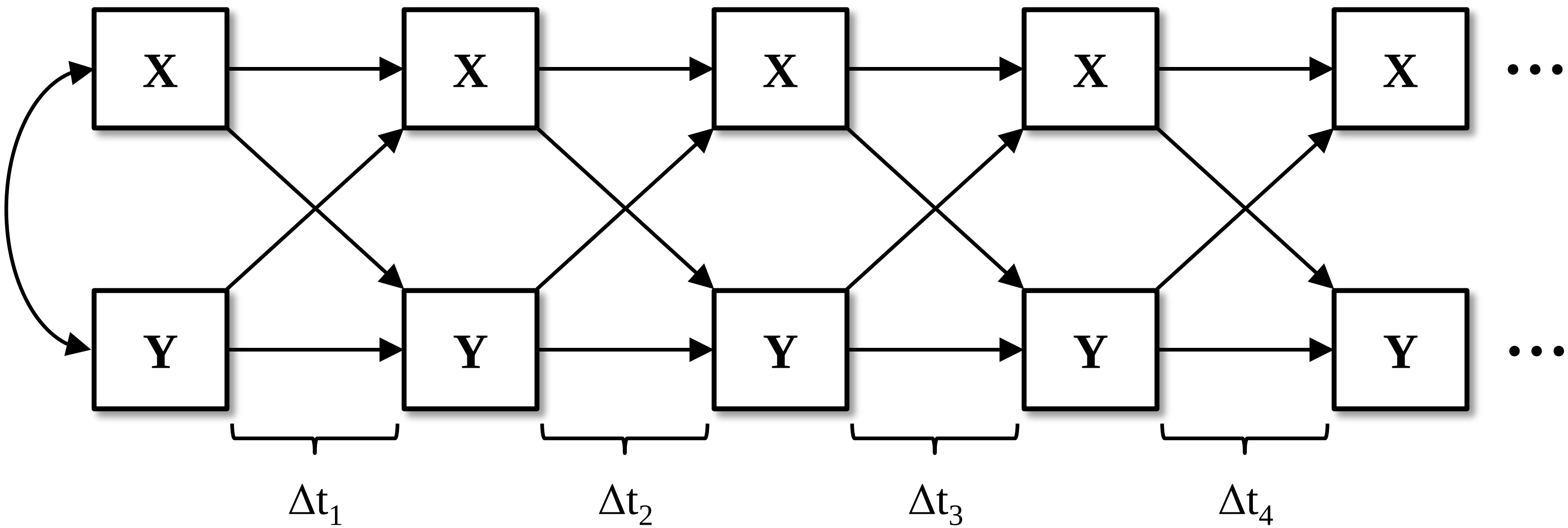
e.g., For two variables:

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \begin{bmatrix} w_{(x)t} \\ w_{(y)t} \end{bmatrix}$$

Using this model, one can assess the stability of constructs as well as the relationships between constructs over time



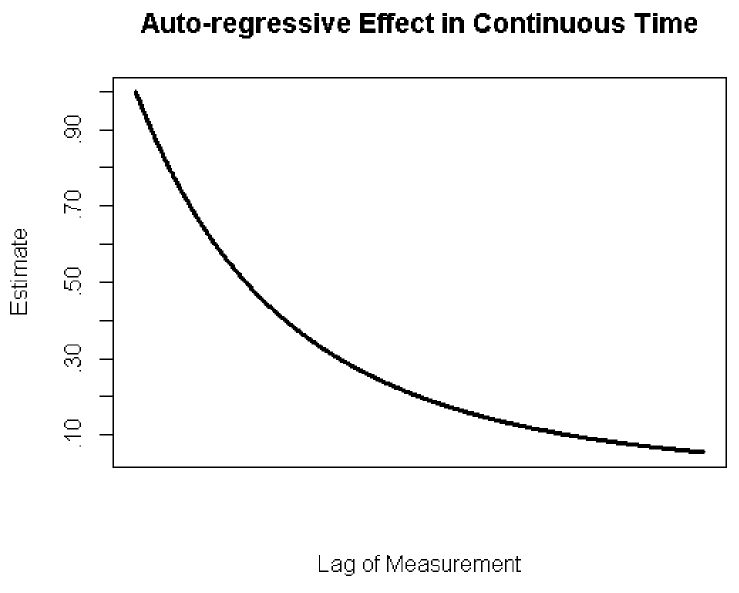
A drawback of this approach is that the estimates depend on the time elapsed between observation intervals

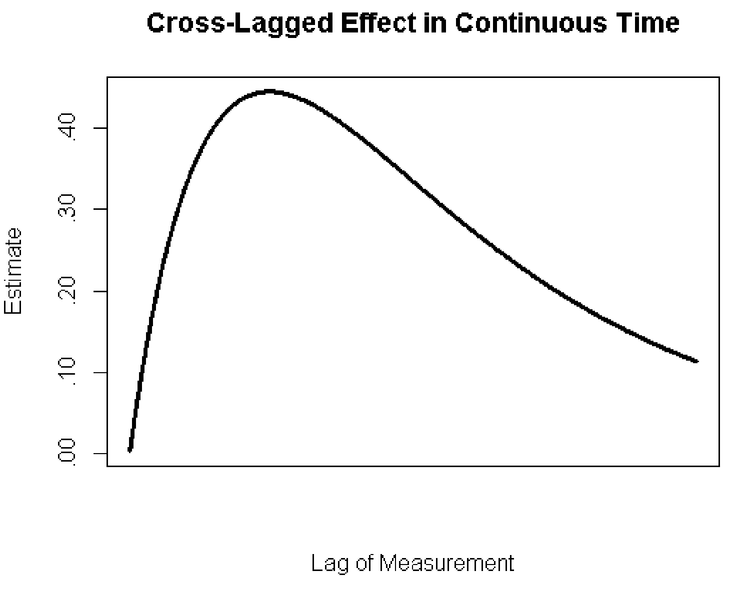


Therefore, the model is more appropriately expressed like this. Notice that the discrete-time parameter matrix $\mathbf{A}$ depends on the observation interval Δt.

$$\mathbf{x}(t) = \mathbf{A}(\Delta t)\mathbf{x}(t - \Delta t) + \mathbf{w}(\Delta t)$$

For a first-order (AR1) model, the effects observed in the discrete-time CLPM follow a specific pattern in continuous-time

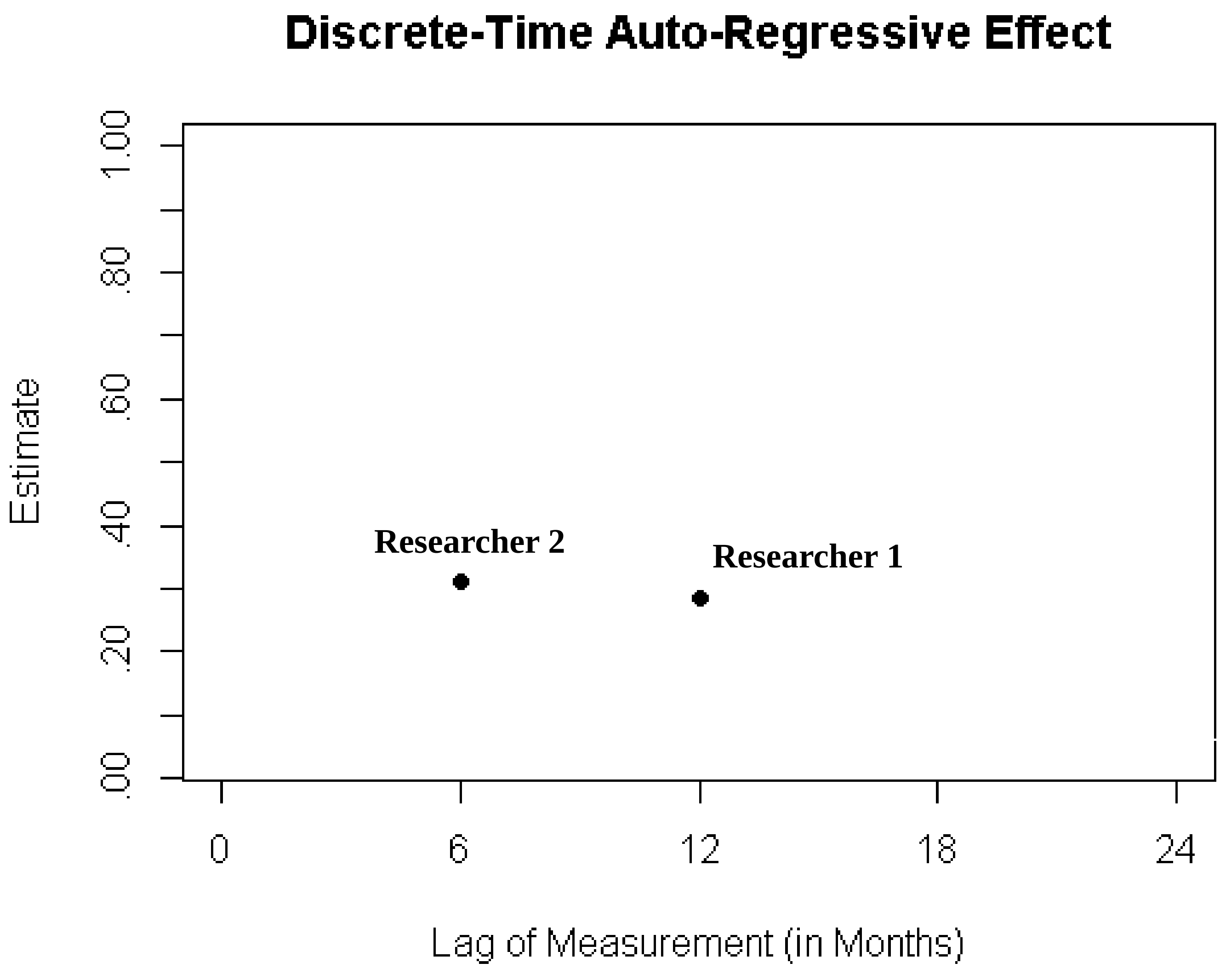


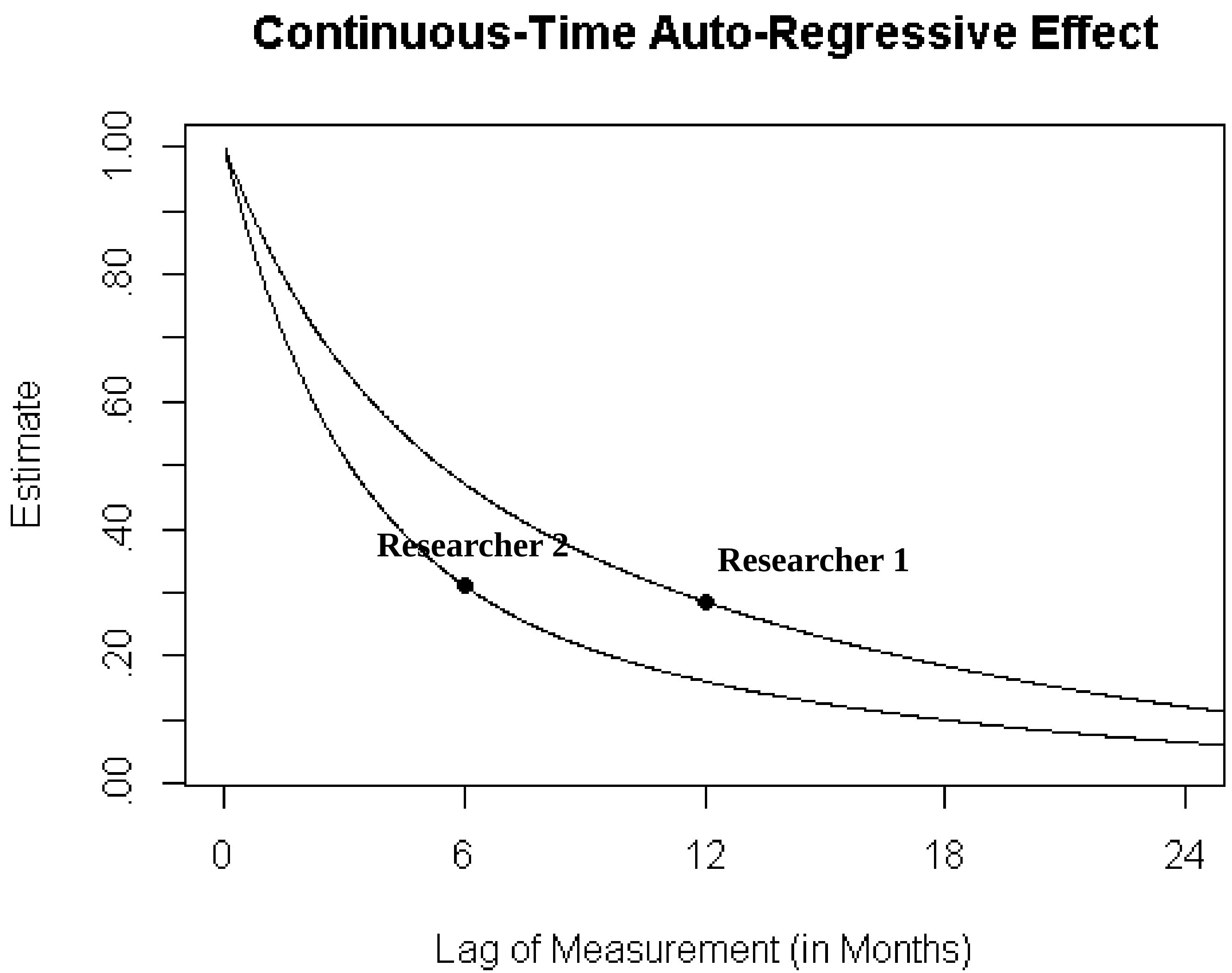


Due to the dependence of the discrete-time model parameters on the lag of measurement, the cross-lagged panel model is in some cases limited

Imagine a study in which two researchers wish to study the stability of bullying in two different populations. Researcher 1 investigates bullying in rural areas, and Researcher 2 investigates bullying in urban areas

Researcher 1 surveys students at yearly intervals and Researcher 2 surveys students at six-month intervals. The autoregressive coefficient is estimated to be approximately .30 in both studies, leading the researchers to conclude that bullying is equally stable in the two populations

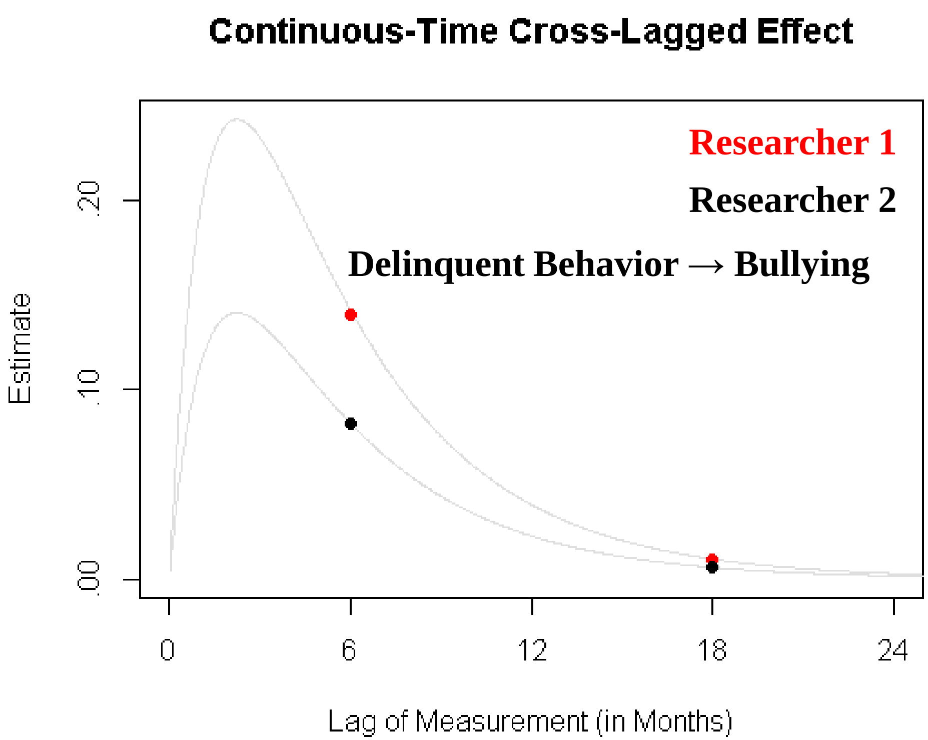

Discrete-Time Auto-Regressive Effect
Estimate
1.00
.80
.60
.40
.20
.00
Researcher 2
Researcher 1
0
6
12
18
24
Lag of Measurement (in Months)


Continuous-Time Auto-Regressive Effect
Estimate
1.00
.80
.60
.40
.20
.00
Researcher 2
Researcher 1
0
6
12
18
24
Lag of Measurement (in Months)

Equal observation intervals are not always a panacea

Suppose that both researchers examined the effect of delinquent behavior on bullying (and vice versa). The researcher used equal observation intervals.

Even with equal intervals, the comparison between the two studies will depend on the measurement lag chosen



Therefore, there may be some advantage to examine the data-generating process in continuous time

Differential equations can be used to model processes in continuous time. The continuous-time equivalent model of the CLPM is:

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{A}\mathbf{x}(t)$$

In this differential equation, the rate of change in the state of the system (i.e., the variables in vector $\mathbf{x}$) at time t is proportional to the current state of the system

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{A}\mathbf{x}(t)$$

To estimate the continuous time parameters in drift matrix $\mathbf{A}$ with discrete observations, we must connect the parameters in $\mathbf{A}$ to the discrete time parameters in $\mathbf{A}(\Delta t_i)$. The relationship between the two parameter matrices is:

$$\mathbf{A}(\Delta t_i) = e^{\mathbf{A}\Delta t_i}$$

This relationship forms the basis of the exact discrete model (EDM)

The EDM is an exact solution to the (expanded) stochastic differential equation:

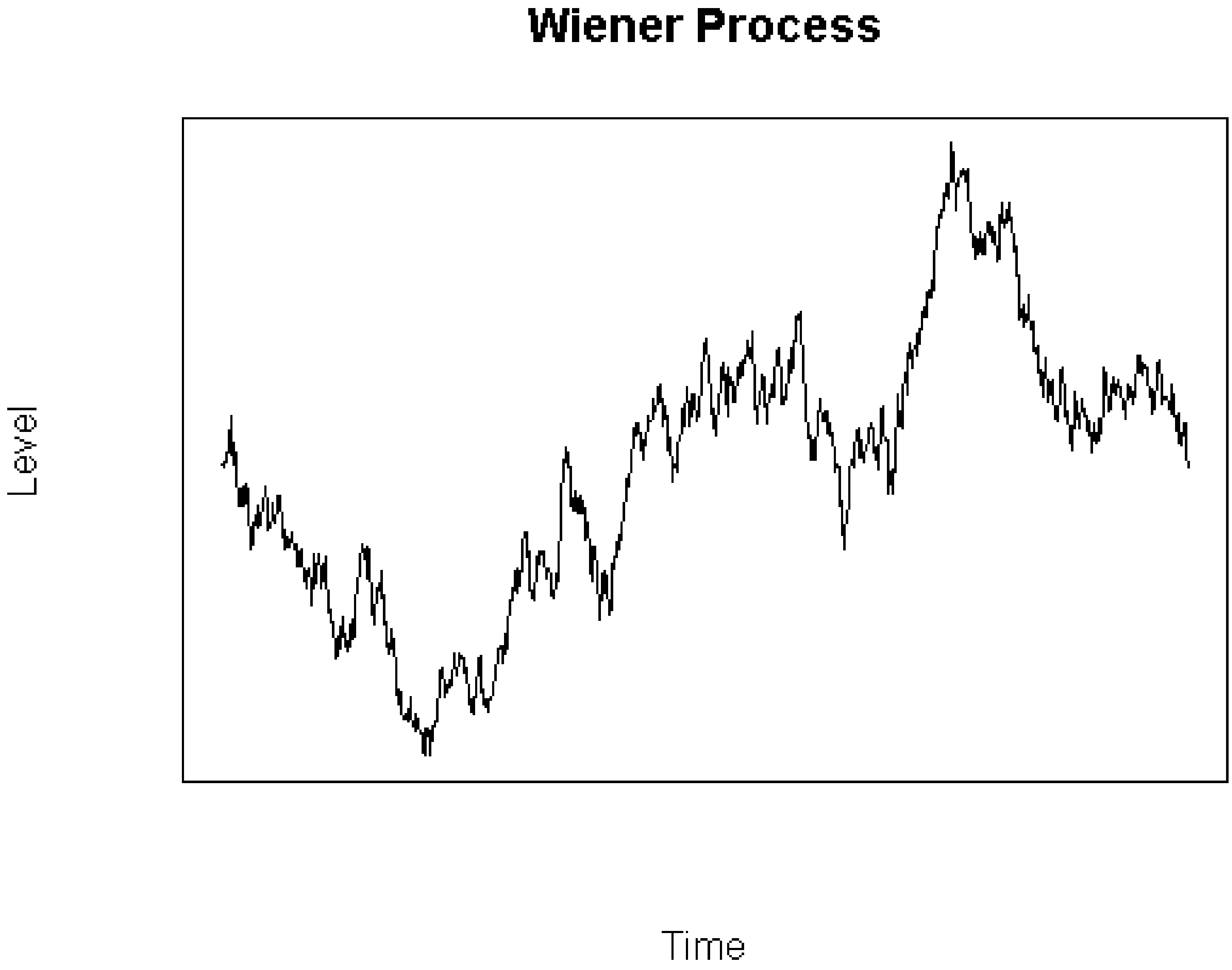
$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{A}\mathbf{x}(t) + \mathbf{b} + \mathbf{G}\frac{d\mathbf{W}(t)}{dt}$$

In this expanded version, we have added a vector $\mathbf{b}$ that contains continuous time intercepts. These parameters allow non-zero mean trajectories for the variables in $\mathbf{x}(t)$

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{A}\mathbf{x}(t) + \mathbf{b} + \mathbf{G}\frac{d\mathbf{W}(t)}{dt}$$

Also, we have added a continuous time error process. $\mathbf{W}(t)$ is a Weiner process (i.e., a random walk) that is "differentiated" and scaled by the matrix $\mathbf{G}$

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{A}\mathbf{x}(t) + \mathbf{b} + \mathbf{G}\frac{d\mathbf{W}(t)}{dt}$$


Wiener Process
Level
Time

The solution to the stochastic differential equation (i.e., the EDM) is:

$$\mathbf{x}(t) = \mathrm{e}^{\mathbf{A}(t-t_0)}\,\mathbf{x}(t_0) + \mathbf{A}^{-1}[\mathrm{e}^{\mathbf{A}(t-t_0)} - \mathbf{I}]\mathbf{b} + \int_{t_0}^{t} \mathrm{e}^{\mathbf{A}(t-s)}\mathbf{G}\,\mathrm{d}\mathbf{W}(s)$$

$$\text{with}\ \mathrm{cov}\left[\int_{t_0}^{t} \mathrm{e}^{\mathbf{A}(t-s)}\mathbf{G}\,\mathrm{d}\mathbf{W}(s)\right] = irow\left\{\mathbf{A}_{\#}^{-1}\left[\mathrm{e}^{\mathbf{A}_{\#}(t-t_0)} - \mathbf{I}\right] row\mathbf{Q}\right\}$$

$$\text{and}\ \mathbf{A}_{\#}^{-1} = \mathbf{A}\otimes\mathbf{I} + \mathbf{I}\otimes\mathbf{A},\ \ \mathbf{Q} = \mathbf{G}\mathbf{G}'$$

Oud & Jansen (2000) showed that the EDM can be estimated within the structural equation modeling framework

In the following example, we use the CT_SEM.R program (Voelke, Oud, Davidov, & Schmidt, 2012) which uses the OpenMx SEM package. The program code is available at http://dx.doi.org/10.1037/a0027543.supp

The example uses data from a RCT of the Youth Matters anti-bullying program (Jenson & Dieterich 2007)

The trial was conducted over two academic years between 2003 and 2005

Data are from the control group (N = 461)

We will be looking at the mutual influence of bullying and anti-social attitudes over four observation periods

The EDM can accommodate latent variables via a measurement equation:

$$\mathbf{y} = \boldsymbol{\upsilon} + \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon} \text{ with } \mathrm{cov}(\boldsymbol{\varepsilon}) = \boldsymbol{\Theta}$$

but in this example we will be using observed variables.

To use the CT_SEM.R program, OpenMx must be installed. You can install it by running the following command in the R console:

```
source('http://openmx.psyc.virginia.edu/getOpenMx.R')
```

The CT_SEM program requires the user to make changes to the syntax. These can be broken down into two parts: (1) Data processing and (2) Providing starting values

In the following slides, required user changes are highlighted in orange

For data processing, the user must provide:

1. The filepath of the folder in which the data file is saved
2. The datafile name
3. The total number of construct indicators
4. The total number of constructs

```
setwd("[filepath]")
require(OpenMx)
input_file <- "[filename].dat"
n.Manifest <- 2
n.Latent <- 2
```

For data processing, the user must provide:

5. The value or character used to indicate missing values
6. The columns of the dataset to be used in the analysis

```
#-------------------PREPROCESSING-------------------#
Data <- read.table(file = input_file, header = FALSE)
is.na(data)=data==999
Data <- data[,c(1,2,3,4,5,6,7,8)]
...
#-------------------------------------------------#
```

The indicators for the constructs (say X and Y) measured T times must be in a specific order:

e.g., For a single indicator per construct:

Time 1		**Time 2**		**Time 3**		…	**Time T**	
x	y	x	y	x	y		x	y

e.g., For two indicators per construct:

Time 1				**Time 2**				**Time 3**				…	**Time T**			
x1	x2	y1	y2	x1	x2	y1	y2	x1	x2	y1	y2		x1	x2	y1	y2

Once specifications for data processing are set, starting values for all EDM parameters must be entered

Starting values are very important for convergence of the EDM and can be tricky because we are dealing with continuous time parameters

The first matrix we have to specify starting values for is the variance-covariance matrix of the initial observations (i.e., time 0). You can easily obtain starting values by using the cov() function in R on the initial variables

```
PHI1 <- matrix(c(.26, .09,
                 .09, .39
    ),nrow=n.latent, ncol=n.latent, byrow=TRUE)
```

In our example,

```
cov(data[, c(1:2)], use = "pairwise.complete.obs")
           V1         V2
V1 0.26457634 0.08876605
V2 0.08876605 0.39366107
```

```
PHI1 <- matrix(c(.26, .09,
                 .09, .39
     ),nrow=n.latent, ncol=n.latent, byrow=TRUE)
```

In a similar manner, we next provide starting values for the initial variable means. In our example,

```
colMeans(data[, c(1:2)], na.rm = TRUE)
      V1        V2
1.298200 1.319289
```

```
latentM1 <- matrix(c(1.30, 1.32
        ),nrow=n.latent, ncol=1, byrow=TRUE)
```

The next three matrices provide starting values for the measurement model

$$\mathbf{y} = \boldsymbol{\upsilon} + \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon} \text{ with } \operatorname{cov}(\boldsymbol{\varepsilon}) = \boldsymbol{\Theta}$$

In our example we only have observed variables. Thus,

```
manifestM <- matrix(c(0, 0
          ),nrow=n.manifest, ncol=1, byrow=TRUE)

LAMBDA <- matrix(c(1, 0,
                   0, 1
       ),nrow=n.manifest, ncol=n.latent,byrow=TRUE)

THETA <- matrix(c(0, 0,
                  0, 0
      ),nrow=n.manifest, ncol=n.manifest)
```

Next, we have the drift matrix $\mathbf{A}$ with the continuous time auto- and cross-effects

Values on the diagonal of the matrix (auto-effects) are expected to range between $-\infty$ and 0

```
DRIFT <- matrix(c(-0.17, 0.05,
                  0.12,-0.23
      ),n.latent,n.latent)
```

Values on the off-diagonal of the matrix (cross-effects) can range between -∞ and ∞

The directions of the cross-effects are the same as in discrete time

```
DRIFT <- matrix(c(-0.17, 0.05,
                   0.12,-0.23
      ),n.latent,n.latent)
```

One simple method to find starting values is to take the log of the discrete time matrix $\mathbf{A}(\Delta t_i)$ and divide by the lag (here using the expm package)

```
require(expm)
dtime <- matrix(c(.42, .09,
                  .22, .29),
                nrow = 2, ncol = 2, byrow = TRUE)
logm(dtime) / 5.7
           [,1]        [,2]
[1,] -0.1658094  0.04767878
[2,]  0.1165481 -0.23467875
```

Next, we have the continuous-time intercept vector $\mathbf{b}$ with the continuous-time auto- and cross-effects

These can range between -∞ and ∞. Starting at 0 and working outward is helpful.

```
CINT <- matrix(c(.20, .20
     ),ncol=n.latent, nrow=1, byrow=TRUE)
```

Finally, we have the diffusion matrix $\mathbf{Q}$ with the continuous-time residual variances and covariances

Values in this matrix will range from 0 to ∞

```
Q <- matrix(c(.11, .01,
              .01, .21
  ),n.latent, n.latent)
```

Notice we have allowed correlation between the residuals to be a free parameter

Values for this matrix will typically be small but it will depend on the scale

```
Q <- matrix(c(.11, .01,
              .01, .21
  ),n.latent, n.latent)
```

The last piece of code that needs to be changed is the Δt_i vector that contains the length between successive intervals. It is important to choose an appropriate scale for time in this vector

```
delta_t <- c(4,9,4)
```

In our example, the measurement lags were:

Δt_1 = 4 months (between times 0 and 1)

Δt_2 = 9 months (between times 1 and 2)

Δt_3 = 4 months (between times 2 and 3)

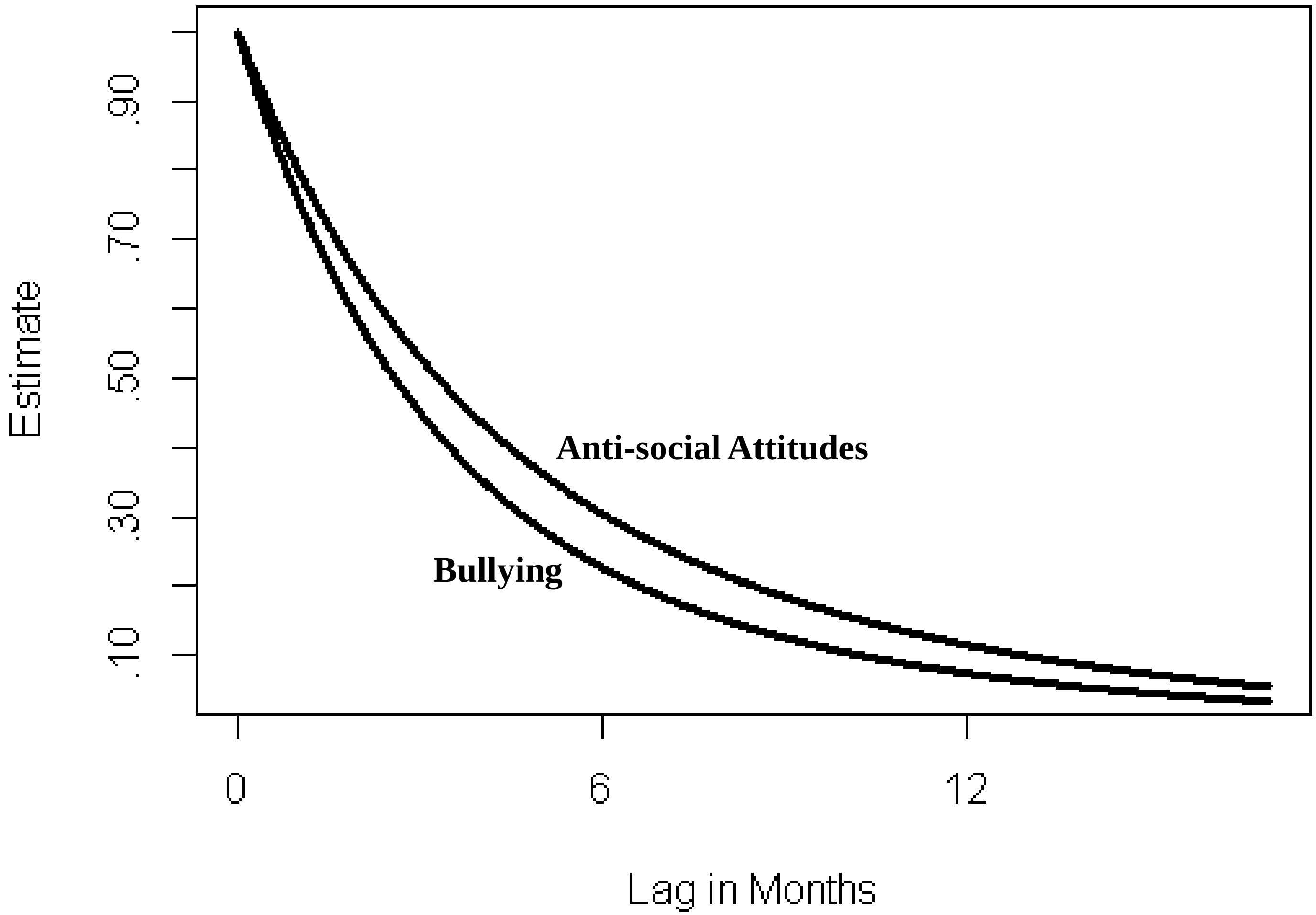
```
delta_t <- c(4,9,4)
```

The rest of the code you can run as is without changing. The results of the analysis will be saved in an object called `fit`

```
#-------DO NOT CHANGE ANYTHING BELOW THIS LINE-------#
...
fit=mxRun(model2)
summary(fit)
```

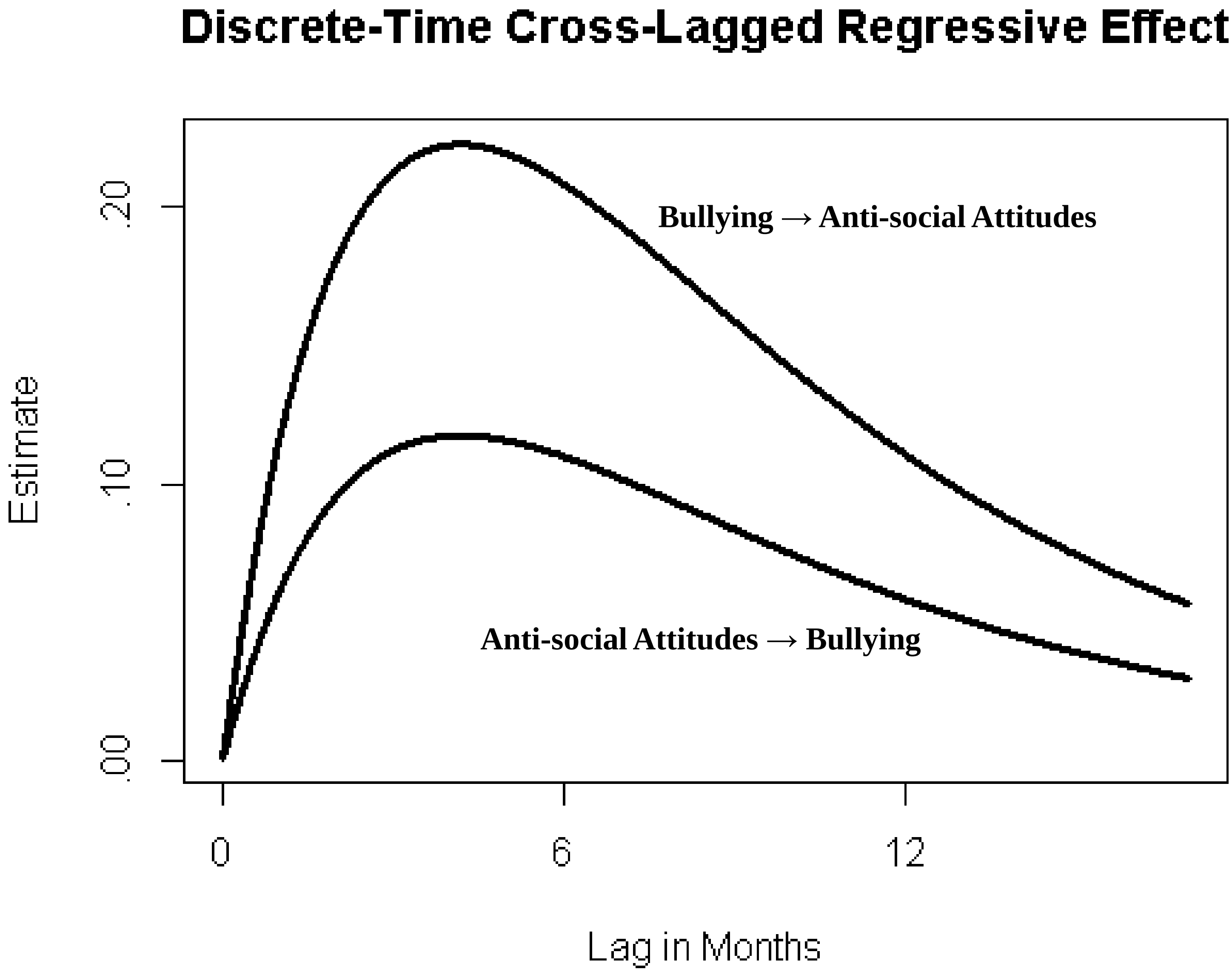
Parameter	Estimate	SE	95% L.B.	95% U.B.
φ_{ASA}	.27	.01	.24	.29
φ_{BUL}	.09	.01	.07	.11
$\varphi_{ASA/BUL}$	.40	.02	.36	.43
μ_{ASA}	1.30	.02	1.27	1.33
$\mu_{BULLYING}$	1.32	.02	1.29	1.36
a_{ASA}	-.23	.02	-.26	-.20
$a_{BULLYING}$	-.28	.02	-.32	-.24
$a_{ASA \rightarrow BUL}$	.15	.02	.10	.20
$a_{BUL \rightarrow ASA}$	.08	.01	.05	.11
b_{ASA}	.21	.02	.18	.24
$b_{BULLYING}$	.18	.02	.14	.23
q_{ASA}	.11	.01	.09	.12
$q_{BULLYING}$	.21	.01	.19	.24
$q_{ASA/BUL}$	.02	.01	.00	.03

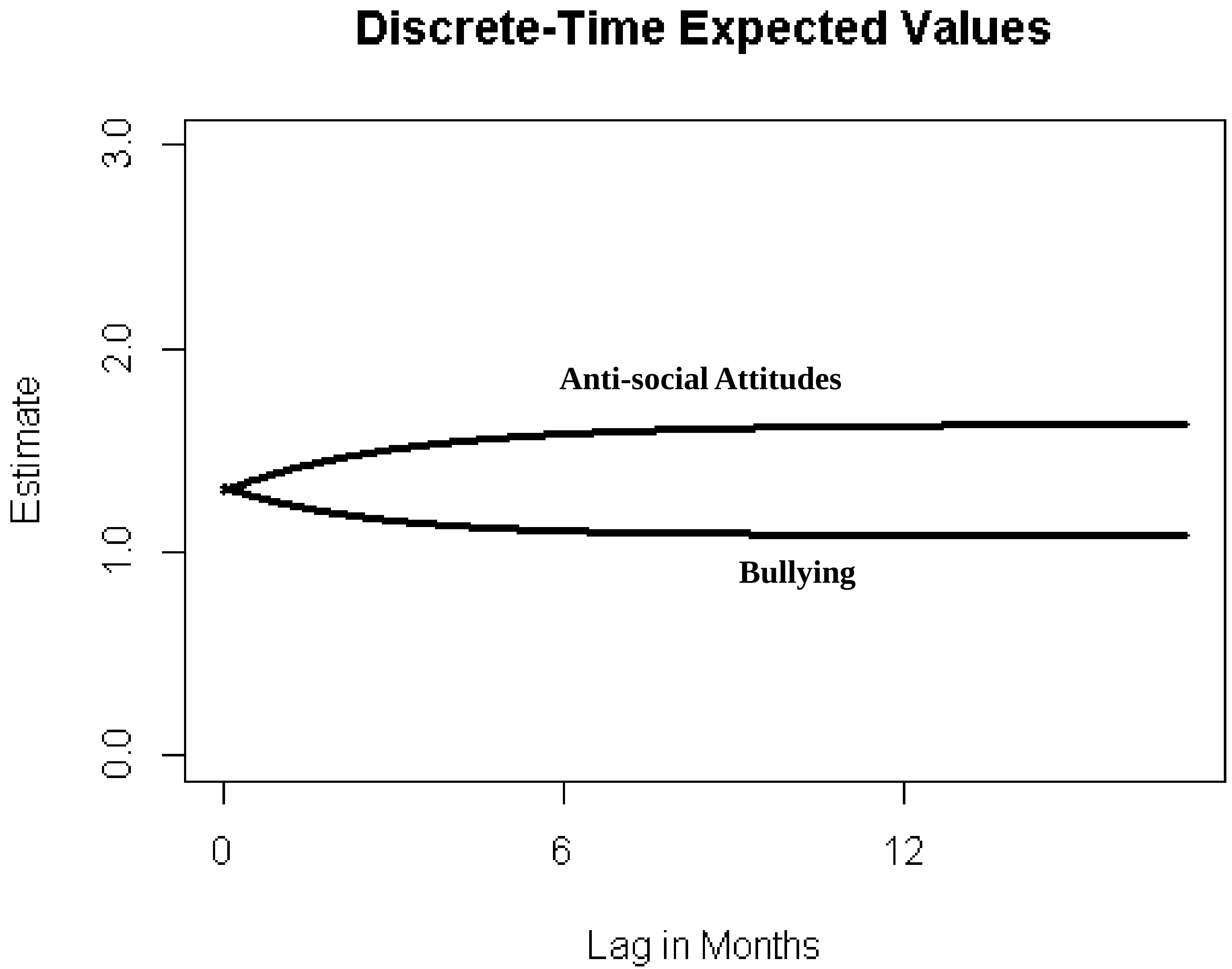
Discrete-Time Auto-Regressive Effect



$$\mathbf{A}(\Delta t_i) = e^{\mathbf{A}\Delta t_i}$$

```
Acon <- matrix(c(-.2281, .0789,
                 .1495,-.2836), nrow = 2)
require(expm)
output <- vector(mode = "list", length = 1700)
for (i in 1:1700) {
    Lag <- i / 100
    Adelta <- expm(Acon * Lag)
    output[[i]] <- Adelta
}
```


Discrete-Time Cross-Lagged Regressive Effect
Bullying → Anti-social Attitudes
Anti-social Attitudes → Bullying
Estimate
.00
.10
.20
0
6
12
Lag in Months


Discrete-Time Expected Values
Estimate
3.0
2.0
1.0
0.0
Anti-social Attitudes
Bullying
0
6
12
Lag in Months

To calculate the mean trends in continuous time, we need to use the following equation:

$$E[\mathbf{x}(t)] = \mathrm{e}^{\mathbf{A}(t-t_0)}\, E[\mathbf{x}(t_0)] + \mathbf{A}^{-1}[\mathrm{e}^{\mathbf{A}(t-t_0)} - \mathbf{I}]\mathbf{b}$$

```
Acon <- matrix(c(-.2281, .0789,
                 .1495,-.2836), nrow = 2)
require(expm)
inmeans <- c(1.30, 1.32)
b <- c(.21, .18)
Iden <- matrix(c(1, 0, 0, 1), ncol = 2, byrow = T)
asa <- vector(mode = "numeric", length = 1700)
bul <- vector(mode = "numeric", length = 1700)
for (i in 1:1700) {
    Lag <- i / 100
    expect <- expm(Acon * Lag) %*% inmeans +
solve(Acon) %*% (expm(Acon * Lag) - Iden) %*% b
    asa[i] <- expect[1]
    bul[i] <- expect[2]
}
```

Several extensions to the EDM are possible:

Oscillating processes (second-order diff. equation)

Individually time-varying intervals

Random effects

On a final note, it is important to point out that the EDM rests on the assumption that the process observed is the same between individuals and across time

Also, the EDM is just a single continuous time model. Many more models may be plausible and have interesting applications in a variety of research situations

Jenson, J. M., & Dieterich, W. A. (2007). Effects of a skills-based prevention program on bullying and bully victimization among elementary school children. *Prevention Science*, *8*, 285-296.

Oud, J. H. L., & Jansen, R. A. R. G. (2000). Continuous time state space modeling of panel data by means of SEM. *Psychometrika*, *65*(2), 199–215.

Voelkle, M. C., Oud, J. H. L., Davidov, E., & Schmidt, P. (2012). An SEM approach to continuous time modeling of panel data: relating authoritarianism and anomia. *Psychological Methods*, *17*, 176–192.