Beyond Linearity, Stability, and Equilibrium: The edm Package for Empirical Dynamic Modeling and Convergent Cross Mapping in Stata

Jinjing Li

NATSEM, University of Canberra

QMNET Seminar Series

Li, J., Zyphur, M., Sugihara, G., & Laub, P. (forthcoming). *Beyond Linearity, Stability, and Equilibrium: The edm Package for Empirical Dynamic Modeling and Convergent Cross Mapping in Stata.* Stata Journal

Outline

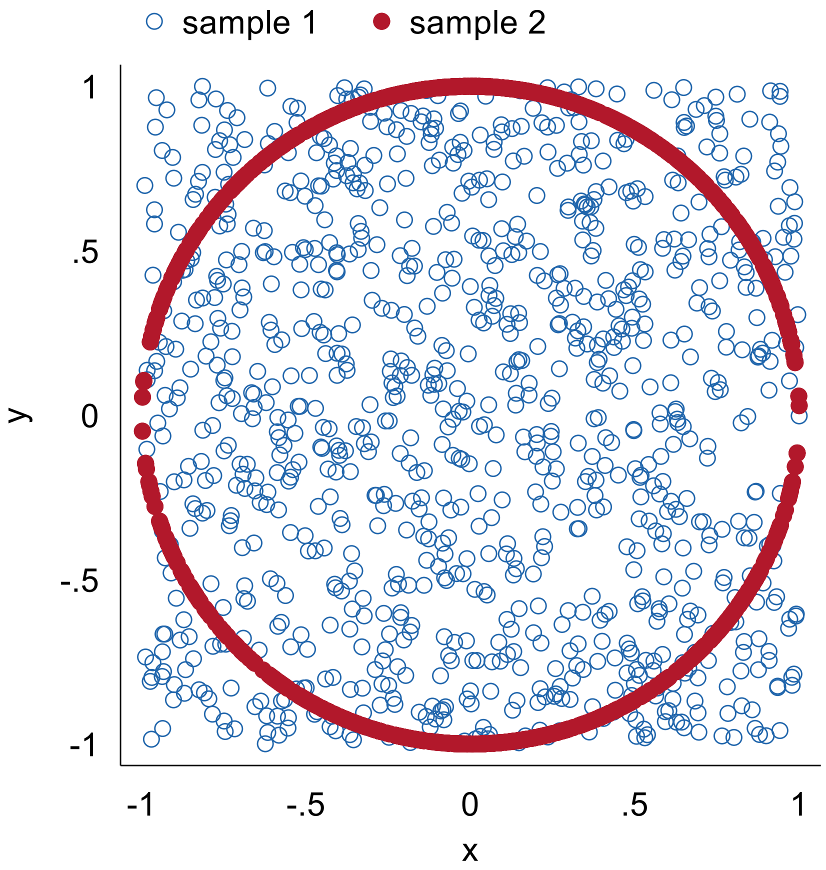
- Causal identifications
- Empirical dynamic modelling
- Implementations in Stata
- Additional features
- Limitations

Causal Identifications

- Causal identifications are fundamental underpinnings in many fields
- Experiments
  - Design, time, cost, ethical concerns
- Using observed data
  - Exploit observed variations
  - Mimic experiments (e.g. PSM, DiD, RD)
  - Incorporate additional information/assumptions (e.g. IV)

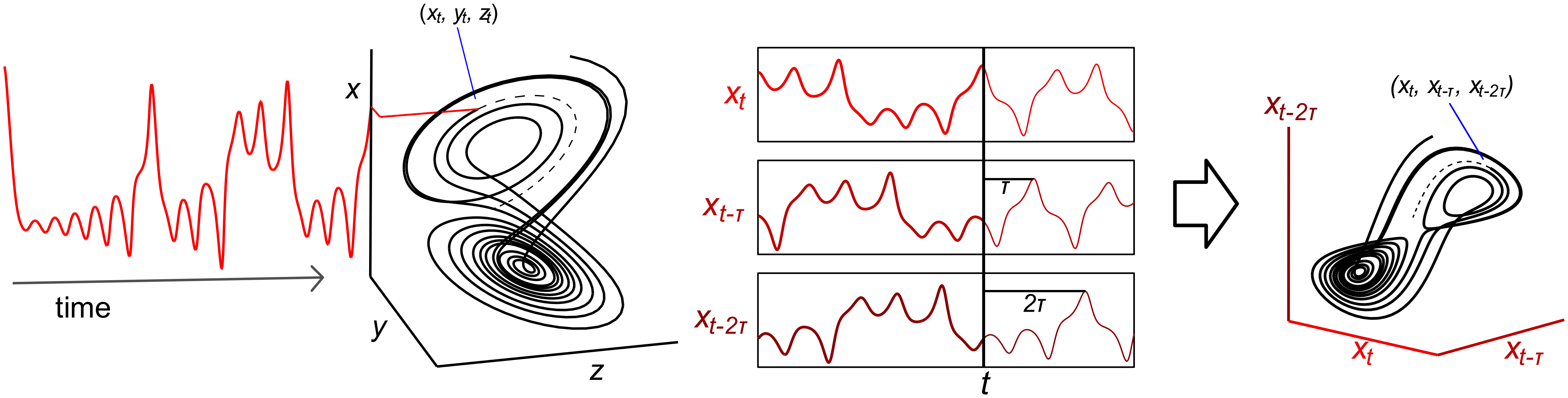
Causal Identifications

- Correlation vs Causation
- Model specification is not known *a priori*
  - Linearity assumption may not be suitable for dynamic complex system
- Granger causality
  - Separability and linearity assumption



Empirical Dynamic Modelling

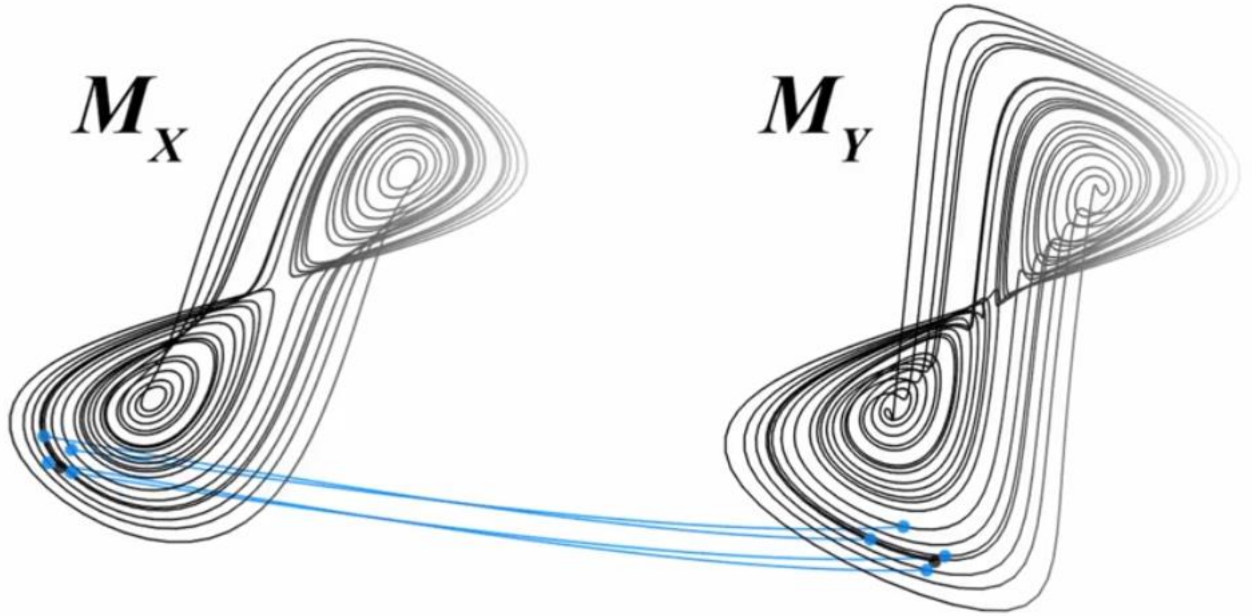
- Based on Takens's theorem



Source: Clark(2015); Ye, Clark, Deyle, & Sugihara (2018)

Convergent cross-mapping

- Cross-mapping prediction accuracy reflects the causal direction
- Mapping accuracy improves as density of the manifold increases



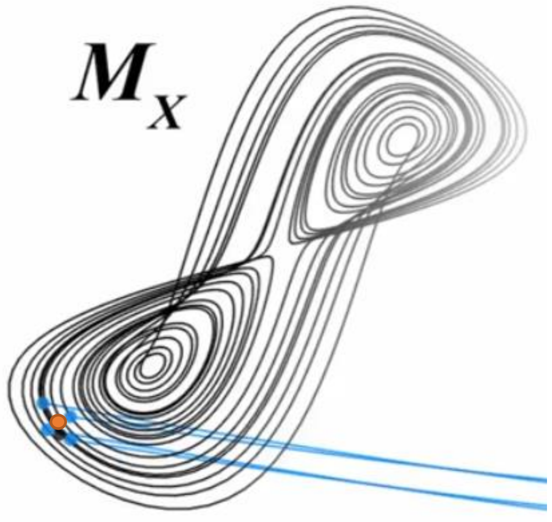
Ye and Sugihara (2012)

Advantages

- Doesn’t need parametric specification
- Can capture complex dynamics
- Reveal causality

Simplex Projection

- Find the nearest k neighbours in the reconstructed manifold
- The prediction is a locally weighted mean
- the weight w_i can be written as $w_i = u_i / \sum_{j=1}^{k} u_j$
- where $u_i = \exp\left(-\theta \frac{\|\boldsymbol{x}_t - \boldsymbol{x}_{t_i}\|}{\|\boldsymbol{x}_t - \boldsymbol{x}_{t_1}\|}\right)$



Ye and Sugihara (2012)

S-map (Sequential locally weighted global linear maps)

- Find the nearest k neighbours in the reconstructed manifold
- Local linear prediction
- SVD solution of $B = AC$
- The vectors are weighted using $w_i = \exp\left(-\theta \frac{\left\|\boldsymbol{x}_t - \boldsymbol{x}_{t_i}\right\|}{\frac{1}{k}\sum_{j=1}^{k}\left\|\boldsymbol{x}_t - \boldsymbol{x}_{t_j}\right\|}\right)$

Evaluate the prediction accuracy

- Use correlation
- Use MAE
- Use other measures

- Compare results from both directions

The edm package in Stata

- Basic syntax

$$\mathbf{edm}\ \begin{matrix}\text{explore}\\ \text{xmap}\end{matrix}\ var1\ var2, \mathrm{e}(int)$$

- The dataset needs to be declared as time-series (tsset) or panel (xtset)
- See **help edm** for other options

The edm package in Stata

- edm explore x

```
. edm explore x

Empirical Dynamic Modelling
Univariate mapping with x and its lag values
-------------------------------------------------------------------
          Actual E            theta                 rho             MAE
-------------------------------------------------------------------
                 2                1              .99908        .0071967
-------------------------------------------------------------------
Note: Random 50/50 split for training and validation data
```

A logistic map example

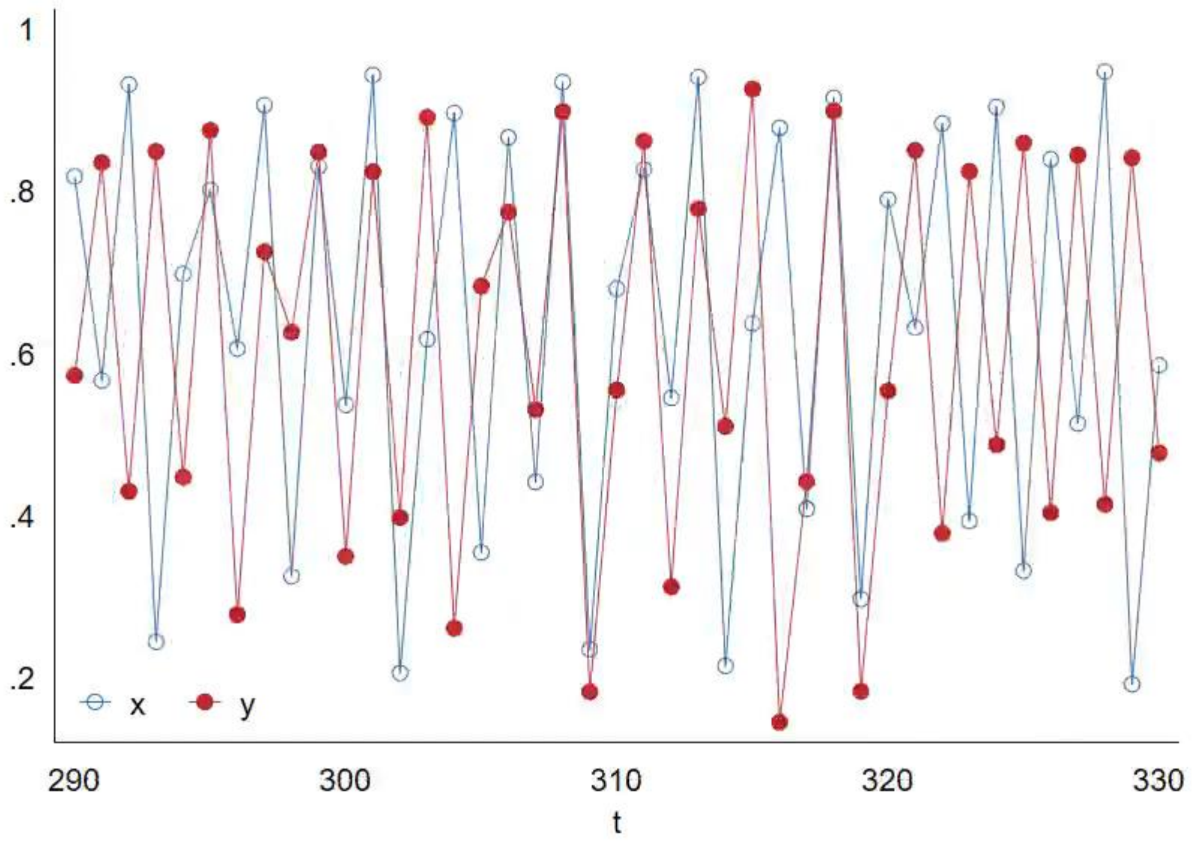
- Setup a chaotic system with two variables

$$\begin{cases} x_t = x_{t-1} * \left(3.79 * (1 - x_{t-1})\right) \\ y_t = y_{t-1} * (3.79 * (1 - y_{t-1}) - 0.20 * x_{t-1}) \end{cases}$$

- First 300 observations burned

```
set obs 500
gen t = _n
tsset t
gen x = 0.2 if _n==1
gen y = 0.3 if _n==1
local r_x 3.79
local r_y 3.79
local beta_xy = 0.0
local beta_yx=0.2
local tau = 1
forvalues i=2/`=_N’ {
    replace x=l.x *(`r_x' *(1-l.x)-
`beta_xy'*l.y) in `i'
    replace y=l.y *(`r_y' *(1-l.y)-
`beta_yx'*l`tau'.x) in `i'
}
keep in 300/450
```

A logistic map example



Correlation between x and y

```
. pwcorr x y, sig

             |        x        y
-------------+------------------
           x |   1.0000
             |
             |
           y |   0.1535   1.0000
             |   0.0599
             |
```

```
. reg x y

      Source |       SS           df       MS      Number of obs   =       151
-------------+----------------------------------   F(1, 149)       =      3.59
       Model |  .214899995         1  .214899995   Prob > F        =    0.0599
    Residual |   8.9102959       149  .059800644   R-squared       =    0.0236
-------------+----------------------------------   Adj R-squared   =    0.0170
       Total |  9.12519589       150  .060834639   Root MSE        =    .24454

------------------------------------------------------------------------------
           x |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           y |   .1680433   .0886453     1.90   0.060    -.0071211    .3432077
       _cons |   .5367766   .0583335     9.20   0.000     .4215088    .6520443
------------------------------------------------------------------------------
```

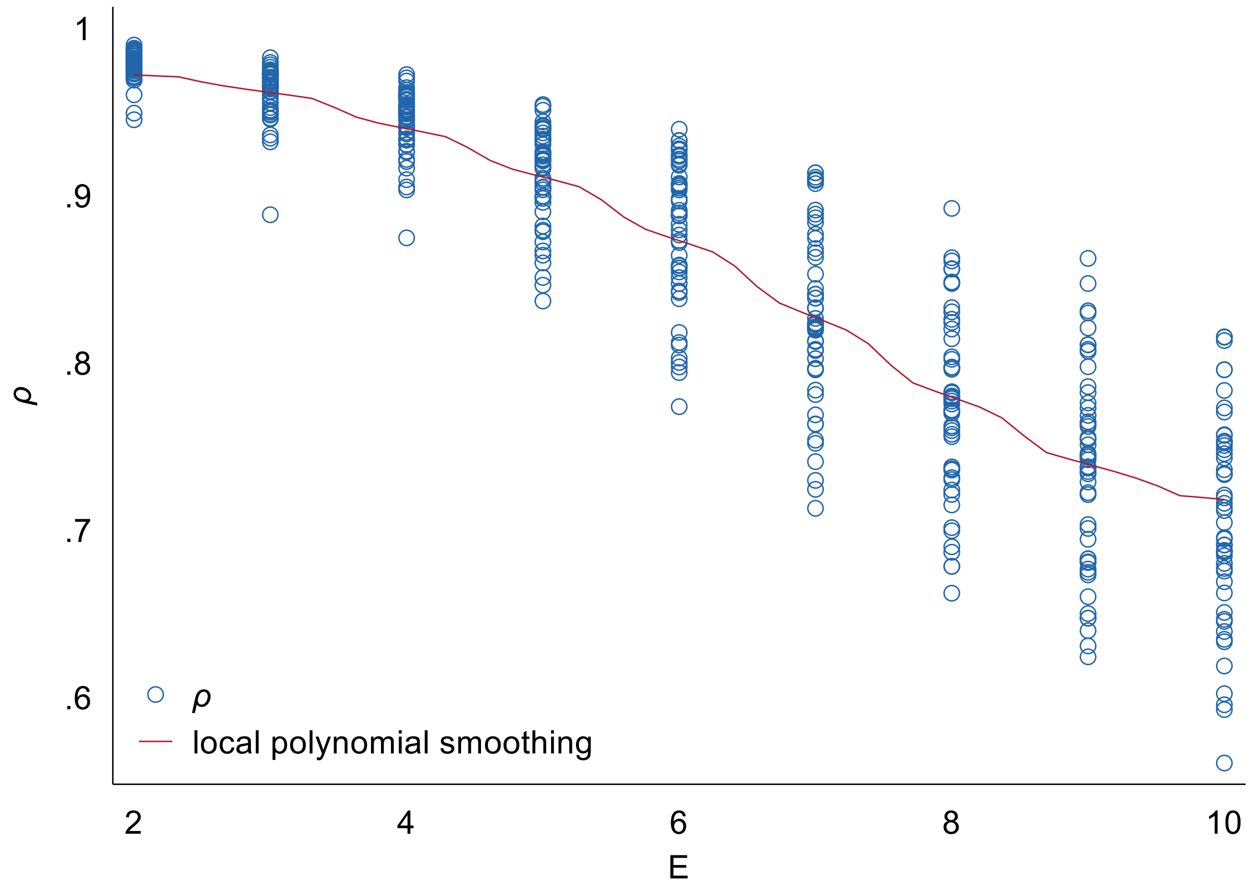
Determine the dimensionality of the system

- Dimensionality is approximated via the prediction accuracy
- Simplex projection with the `explore` subcommand, using the range of dimensions specified in the `e()` option

Determine the dimensionality of the system

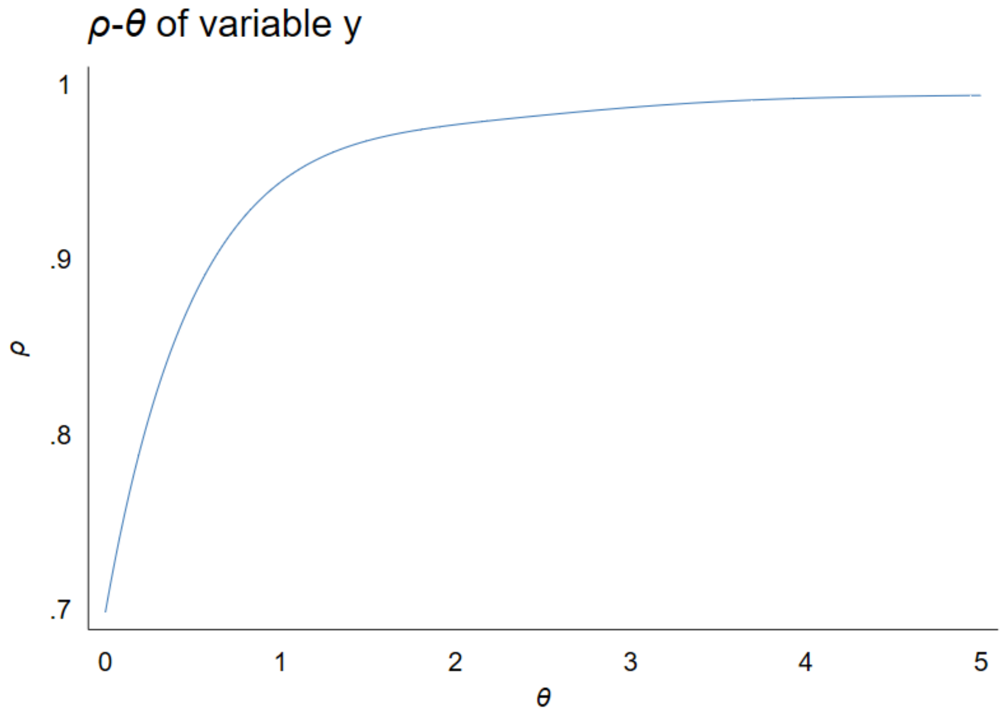
```
. edm explore y, e(2/10) rep(50)

Empirical Dynamic Modelling
Univariate mapping with y and its lag values

----------------------------------------------------------------------
                      --------- rho ---------  --------- MAE ---------
 Actual E    theta          Mean    Std. Dev.         Mean    Std. Dev.
----------------------------------------------------------------------
        2        1        .97818     .0087775      .033184     .0054952
        3        1        .96243      .015995      .042502     .0063758
        4        1        .94326      .019242      .051377     .0067221
        5        1        .91181      .029522      .062658     .0086431
        6        1        .87719      .043446      .072902      .010937
        7        1         .8273      .053334      .085823      .011906
        8        1        .77604      .055847      .099908      .012017
        9        1        .73687      .060581        .1096      .011713
       10        1         .7062      .061469       .11721      .010954
----------------------------------------------------------------------
Note: Results from 50 runs
Note: Random 50/50 split for training and validation data
```

$\rho - E$ plot



Nonlinearity Testing

```
. edm explore y, e(2) algorithm(smap) theta(0(0.01)5) k(-1)

Empirical Dynamic Modelling
Univariate mapping with y and its lag values
--------------------------------------------------------------------
          Actual E            theta              rho              MAE
--------------------------------------------------------------------
                 2                0           .69754           .13328
                 2              .01           .70319           .13226
                 2              .02           .70873           .13125
                 2              .03           .71416           .13024
                 2              .04           .71948           .12924
                 2              .05           .72469           .12825
                 2              .06           .72979           .12727
                 2              .07           .73479           .12629
                 2              .08           .73969           .12532
                 2              .09           .74449           .12436
                 2               .1           .74919           .12341
...
```



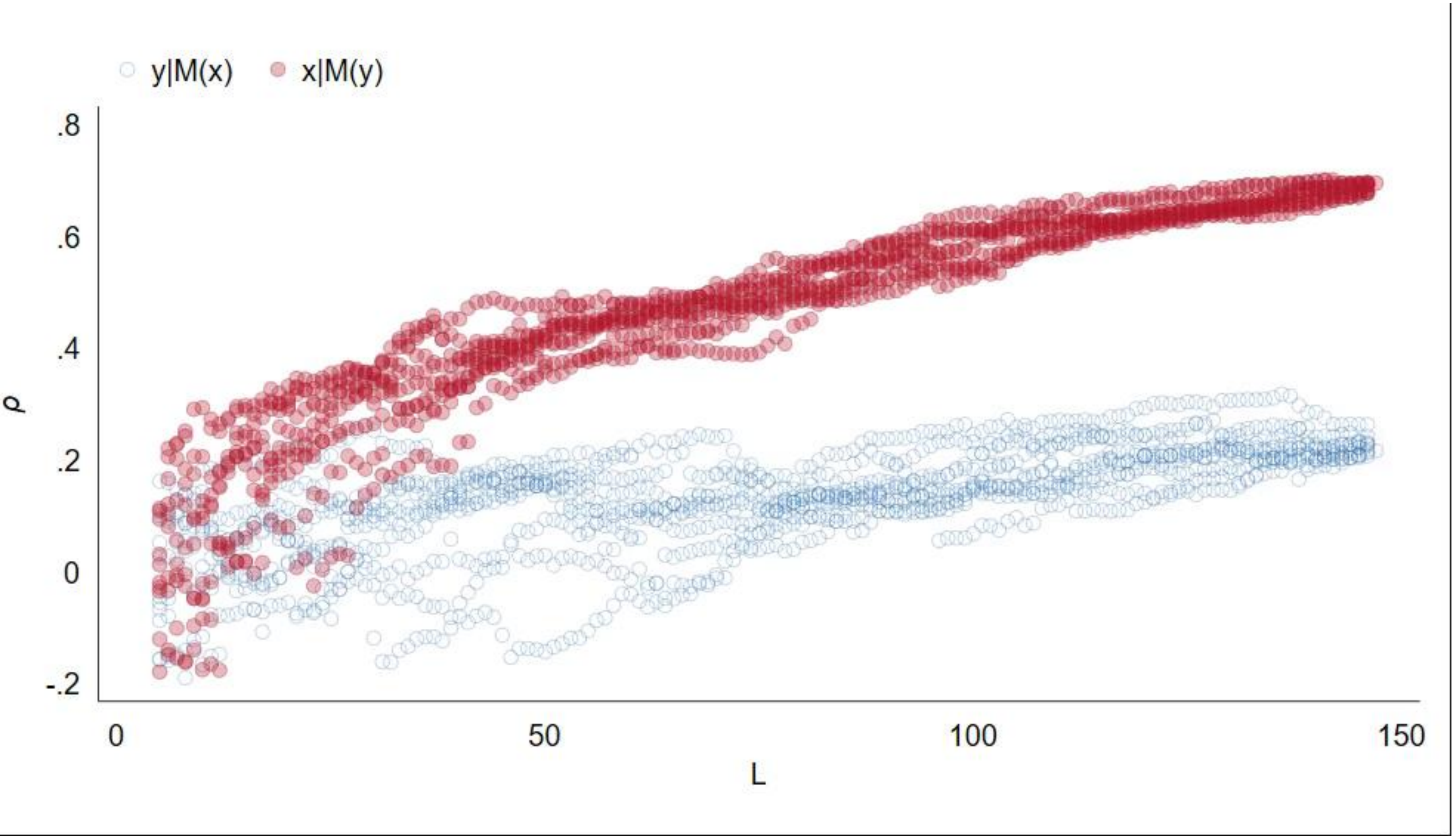
Cross-mapping and causal directions

```
. edm xmap x y, e(2)

Empirical Dynamic Modelling
Convergent Cross-mapping result for variables x and y
--------------------------------------------------------------
      Mapping    Library size             rho              MAE
--------------------------------------------------------------
   y ~ y|M(x)             150          .23019           .19673
   x ~ x|M(y)             150          .69682           .13714
--------------------------------------------------------------
Note: The embedding dimension E is 2
```

Cross-mapping and causal directions

```
edm xmap x y, e(2) rep(10) library(5/150)

Replication progress (20 in total)
....................

Empirical Dynamic Modelling
Convergent Cross-mapping result for variables x and y
---------------------------------------------------------------
      Mapping        Lib size         Mean rho       Std. Dev.
---------------------------------------------------------------
   y ~ y|M(x)               5         -.023008          .10591
   y ~ y|M(x)               6        .00047947          .11217
   y ~ y|M(x)               7          .031598          .10724
   y ~ y|M(x)               8          .030306          .11195
   y ~ y|M(x)               9          .029225           .1057
   y ~ y|M(x)              10          .022898          .10946
```

Cross-mapping and causal directions



Convergence Testing

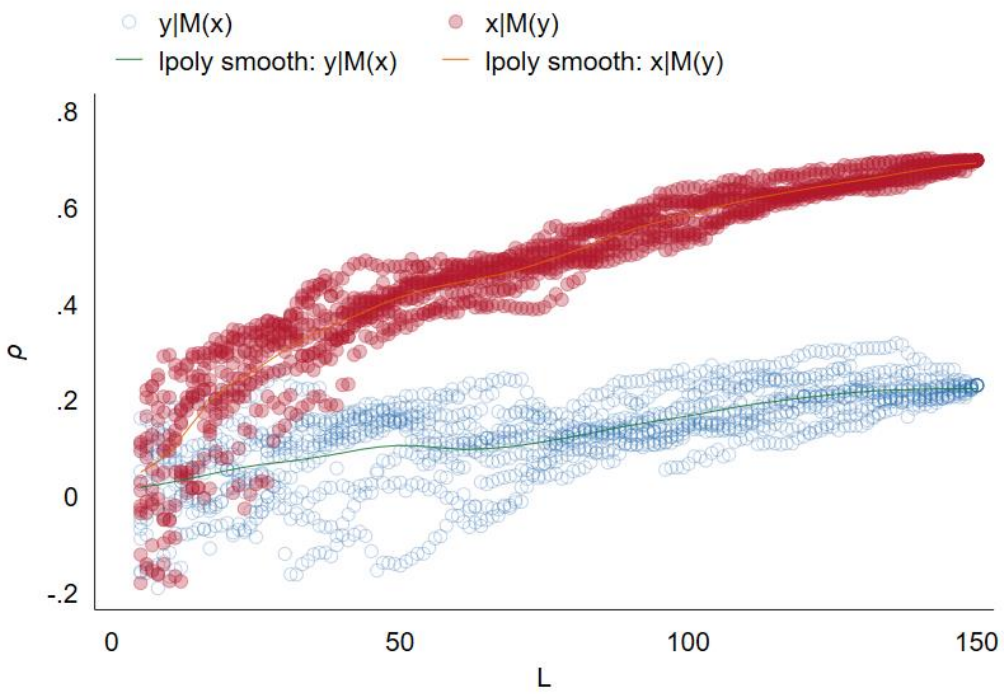
- Parametric fitting
  - Equations with convergence properties such as $\rho_L = \alpha e^{-\gamma L}$
- Hypothesis testing
  - Test $\rho_{100} > \rho_{50}$; $\rho_{X \to Y,L} > \rho_{Y \to X,L}$
- Comparison with a null distribution
  - Randomised timestamp

Convergence Testing

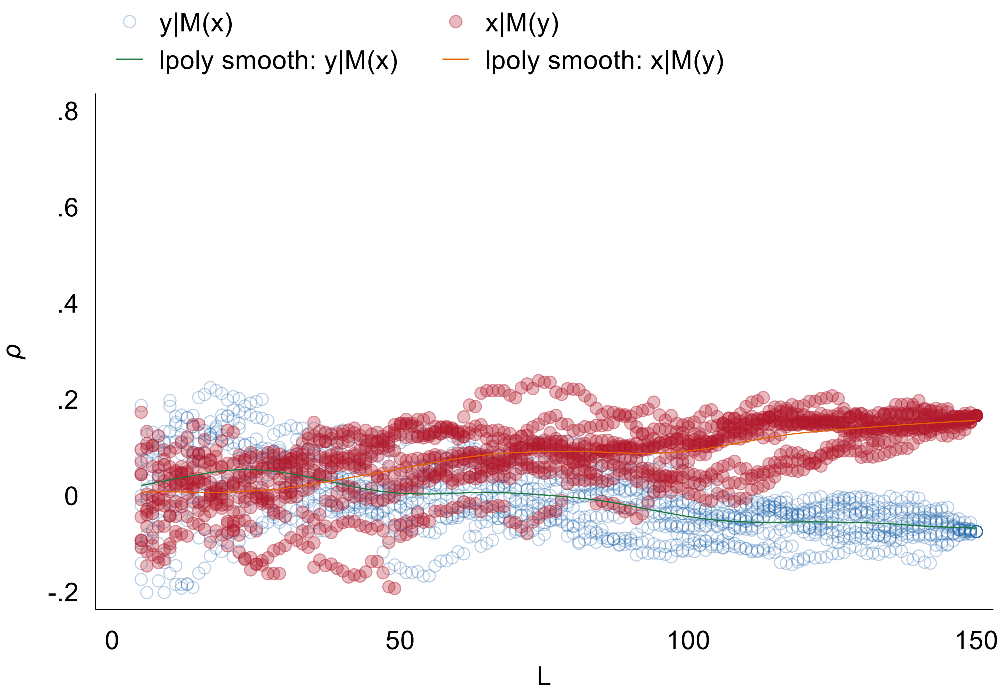
```
. edm xmap x y, library(140) rep(100)
. mat c2= e(xmap_2)
. svmat c2, names(rho140_)

. ttest lib10_yx3 == lib140_yx3, unpaired unequal

Two-sample t test with unequal variances
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
lib10_~3 |     100    .1032279    .0102391    .1023911    .0829113    .1235445
lib140~3 |     100    .6751515    .0012912    .0129119    .6725895    .6777135
---------+--------------------------------------------------------------------
combined |     200    .3891897    .0209145    .2957763    .3479471    .4304323
---------+--------------------------------------------------------------------
    diff |           -.5719236    .0103202               -.5923933   -.5514538
    diff = mean(lib10_yx3) - mean(lib140_yx3)                     t = -55.4178
Ho: diff = 0                     Satterthwaite's degrees of freedom =  102.148

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

Comparison with a null distribution

Original Data



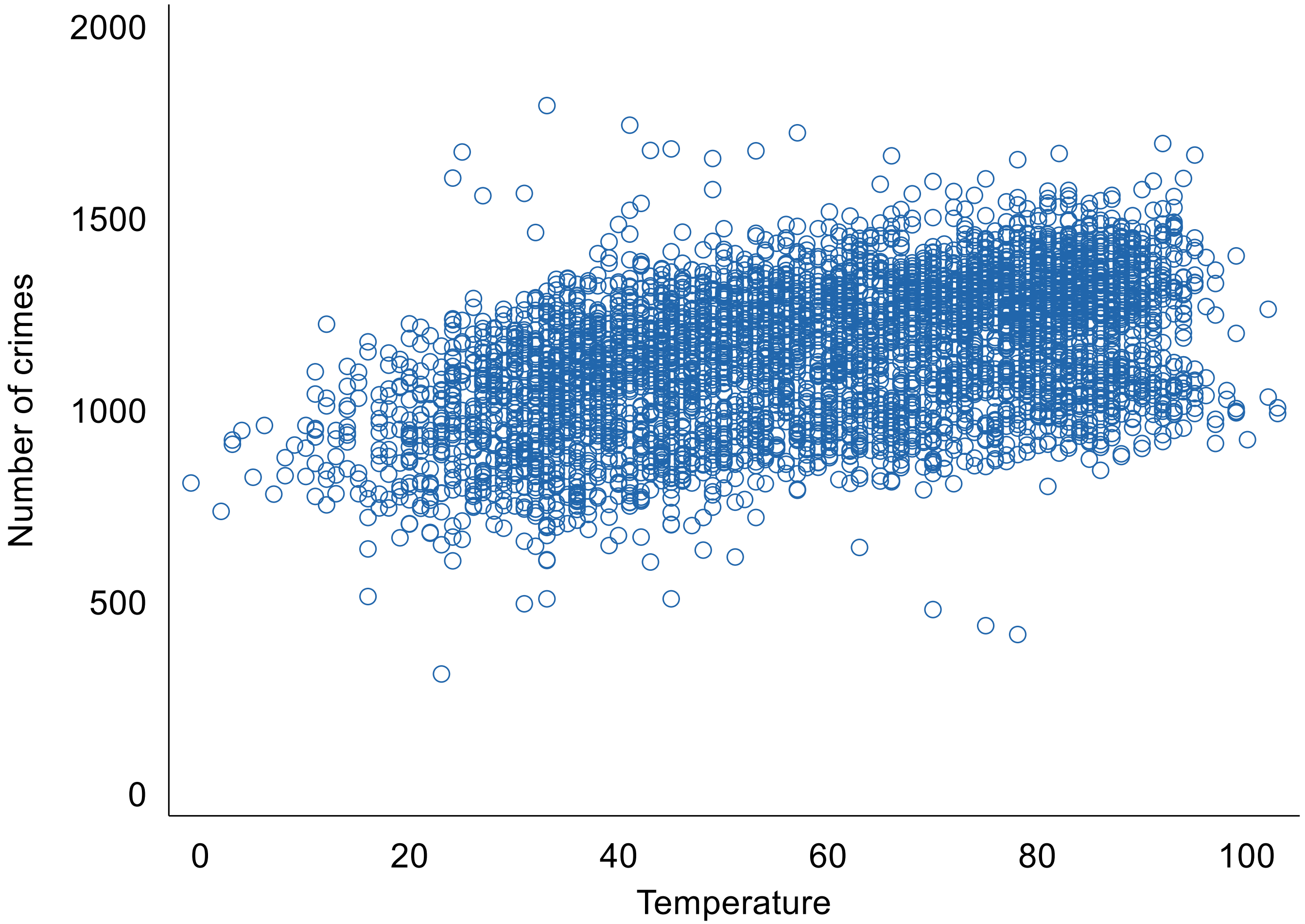
Randomised timestamp



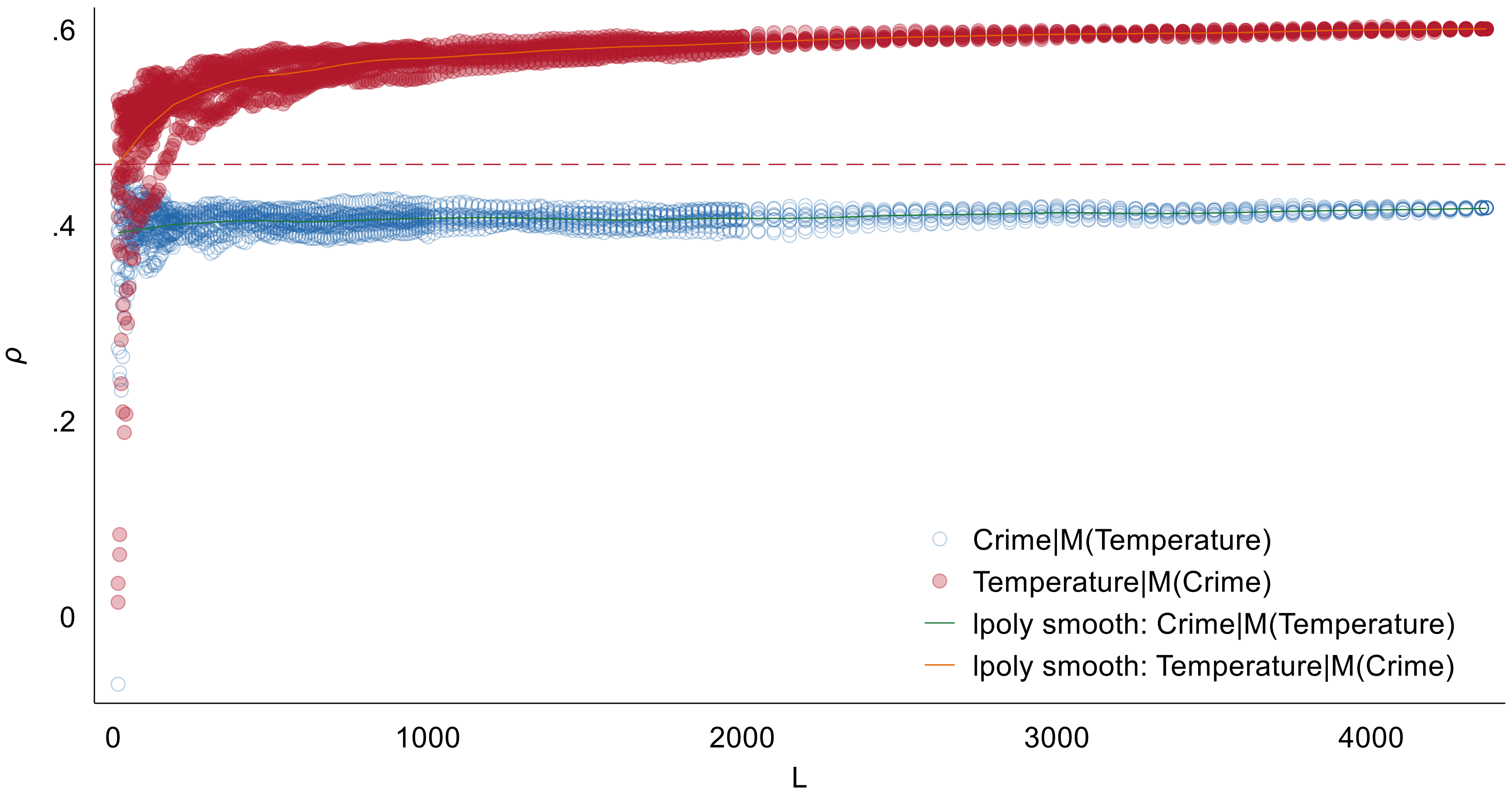
Example: Chicago crime data

- Chicago daily temperature and counts of crime for more than 11 years

- $\rho = 0.46$



Example Chicago crime temperature



Some additional features

- Standardisation of variables
- *k*-fold cross-validations
- Multivariate embedding
- Estimate marginal effect
- EDM with panel data
- Conditional causality
- Time-delayed causality

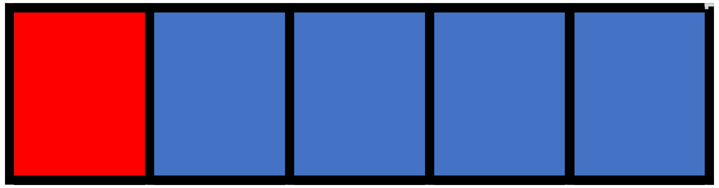
Standardisation of variables

- edm supports the use of time-series operators (e.g. d.x l.x, f.x l.d2.x) and an additional standardisation operator (z.)

- edm explore z.x
- edm xmap z.f.x z.d.y

k-fold cross-validations

```
. edm explore x, cross(5) detail dot(0)

Empirical Dynamic Modelling
Univariate mapping with x and its lag values
--------------------------------------------------------------------
          Actual E                theta               rho          MAE
--------------------------------------------------------------------
                 2                    1            .99956     .0048037
                 2                    1            .99976     .0036994
                 2                    1             .9978     .0088557
                 2                    1            .99914     .0065735
                 2                    1            .99962     .0041738
--------------------------------------------------------------------
Note: Number of neighbours (k) is set to E+1
Note: 5-fold cross validation results reported
```



Multivariate embedding

- It is possible to customize the embedding specification in edm
- edm xmap x y, extra(z l.z l2.z)

Estimate marginal effect using s-map

- edm xmap temp crime, e(7) alg(smap) k(-1) **savesmap(beta)**

```
. desc beta*

              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------------------------------------------------
beta1_b1_rep1   double  %10.0g                temp predicting crime or crime|M(temp) S-map coefficient (rep 1)
beta1_b2_rep1   double  %10.0g                l1.temp predicting crime or crime|M(temp) S-map coefficient (rep 1)
beta1_b3_rep1   double  %10.0g                l2.temp predicting crime or crime|M(temp) S-map coefficient (rep 1)
beta1_b4_rep1   double  %10.0g                l3.temp predicting crime or crime|M(temp) S-map coefficient (rep 1)
beta1_b5_rep1   double  %10.0g                l4.temp predicting crime or crime|M(temp) S-map coefficient (rep 1)
beta1_b6_rep1   double  %10.0g                l5.temp predicting crime or crime|M(temp) S-map coefficient (rep 1)
beta1_b7_rep1   double  %10.0g                l6.temp predicting crime or crime|M(temp) S-map coefficient (rep 1)
beta1_b0_rep1   double  %10.0g                constant in temp predicting crime S-map equation (rep 1)
beta2_b1_rep1   double  %10.0g                crime predicting temp or temp|M(crime) S-map coefficient (rep 1)
```

EDM with panel data

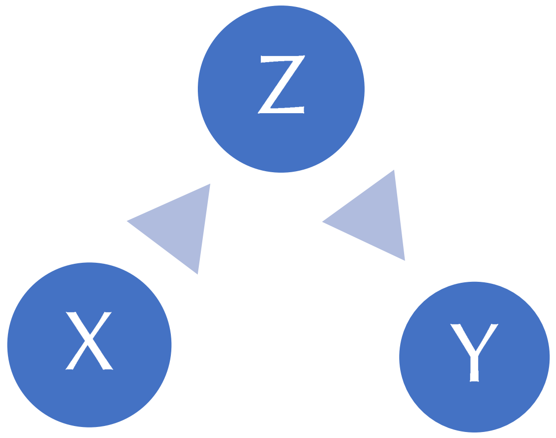
- edm supports panel data (declared by xtset)
- By default it uses the pooled approach
  - Assuming all individual time-series share the same dynamic
- Pre-process data to obtain a fixed-effect like estimator (within estimator)

```
bys id: egen mean_x = mean(x)
gen x_new = x-mean_x
edm explore x
```

Conditional causality

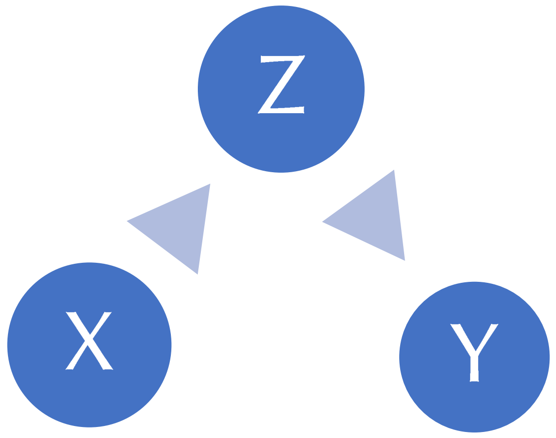
- edm estimates reflect the overall bivariate causality
- It is sometimes useful to estimate the causal direction conditional on an additional variable – partial cross-mapping (Leng et al, 2020)
- Normal cross-map $\rho_{X,\hat{X}|M_Y}$
- Partial cross-map $Pcc(X, \hat{X}|M_Y \mid \hat{X}|M_{\hat{Z}|M_Y})$

where $Pcc(a, b|c) = \frac{\rho_{a,b} - \rho_{a,c}\rho_{b,c}}{\sqrt{(1-\rho_{a,c}^2)(1-\rho_{b,c}^2)}}$



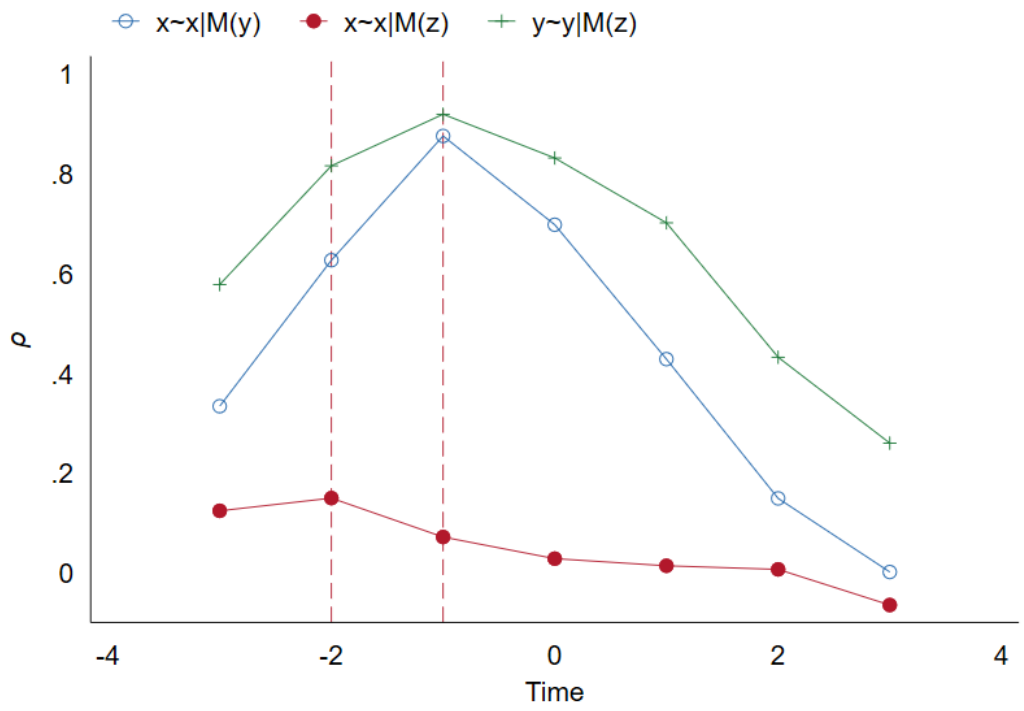
Conditional causality

- edm xmap y x, oneway predict(x_my)
- edm xmap y z, oneway predict(z_my)
- edm xmap z_my x, oneway predict(x_mz_my)
- cor x x_my
- pcorr x x_my x_mz_my



Time-delayed causality

- The edm command supports such reverse- and forward-lagged analyses with minimal input changes by relying on time-series operators (prefixes l. and f.).



Limitations

- Bivariate overall causality
- Hyperparameters
- Estimations might be slow for datasets with large N*T and large E
  - Multi-core CPU/GPU version
  - Cloud-based version
- Missing values
  - Different solutions
- Different data types / distance measures

References

- Li, J., Zyphur, M., Sugihara, G., & Laub, P. (forthcoming). *Beyond Linearity, Stability, and Equilibrium: The edm Package for Empirical Dynamic Modeling and Convergent Cross Mapping in Stata*. Stata Journal

- Stata package installation

```
ssc install edm
```