Transfer Entropy

A presentation in
the information theory seminar
28.11.2008
Department of Information Technology
University of Turku
Atte Tenkanen

Contents

- Background: IT basics
- Mutual information (M)
- Transfer entropy (TE)
- Some sample applications of TE
- TE applied to music analysis
- Questions...

This presentation is mostly based on thesis by Davis [1].

IT basics

- ”*Information* is a decrease in uncertainty”.
- Intuitively: ”Outcomes which happen less frequently would yield more information about a system.”
- *Time-series* = a sequence of time ordered observations of some system.

IT basics

- $p(x,y)$ = the (joint) probability of events x and y occurring at the same time.
  - assumed to be independent
- $p(x,y)/p(x)=p(y|x)$, 'conditional' probability
  - x and y are *independent* if $p(y|x)=p(y)$
  - -> $\underline{p(x,y)}=p(y|x)p(x)=\underline{p(y)p(x)}$

IT basics

- Information I of a particular event *x* is:

  $I(x) = -log(p(x))$

- Shannon entropy:

$$H = \sum_x p(x) I(x) = -\sum_x p(x) \log_a(p(x))$$

Mutual information

- Shannon entropy for two systems:

$$H_1 = -\sum_{x,y} p(x,y)\log(p(x,y))$$

- If the systems are independent, then

$$H_2 = -\sum_{x,y} p(x,y)\log(p(x)p(y))$$

Mutual information

- $H_2 - H_1$: Output of the two systems as though they were independent as opposed to their 'actual' relationship = Mutual Information.

$$
\begin{aligned}
H_2 - H_1 &= -\sum_{x,y} p(x,y)\log(p(x)p(y)) + \sum_{x,y} p(x,y)\log(p(x,y)) \\
&= \sum_{x,y} p(x,y)\left[\log(p(x,y)) - \log(p(x)p(y))\right] \\
&= \sum_{x,y} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)
\end{aligned}
$$

Mutual information

- As an example, suppose there are two systems 1 and 2.

1: 1110110100110011101101

2: 1010101011011011011001

$$p(0_1) = 8/22 = 0.364 \qquad p(0_1, 0_2) = 3/22 = 0.136$$

$$p(1_1) = 14/22 = 0.636 \qquad p(0_1, 1_2) = 5/22 = 0.227$$

$$p(0_2) = 9/22 = 0.410 \qquad p(1_1, 0_2) = 6/22 = 0.273$$

$$p(1_2) = 13/22 = 0.590 \qquad p(1_1, 1_2) = 8/22 = 0.364$$

Mutual information

- Mutual information

$$\sum_{x,y} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

is calculated as:

$$\begin{aligned}
= \quad & 0.136 \log\left(\frac{0.136}{0.364 \cdot 0.410}\right) + 0.227 \log\left(\frac{0.227}{0.364 \cdot 0.590}\right) \\
& +0.273 \log\left(\frac{0.273}{0.636 \cdot 0.410}\right) + 0.364 \log\left(\frac{0.364}{0.636 \cdot 0.590}\right) \\
= \quad & 6.08 \times 10^{-4}
\end{aligned}$$

```
#################################
## FUNCTION MUTUAL INFORMATION ##
#################################

# INPUT: two time-series vectors X and Y, have to be of the same length.

mi<-function(X,Y)
{
  L1=length(X)
  TPvector=rep(0,L1) # Init.
  for(i in 1:L1)
  {
    TPvector[i]=paste(c(X[i],"i",Y[i]), collapse="") # "addresses"
  }
  TPvector1=table(TPvector)/length(TPvector) # Table of bin-probabilities.
  TPvectorX=table(X)/length(X)
  TPvectorY=table(Y)/length(Y)
  SUMvector=rep(0,length(TPvector1))
  for(n in 1:length(TPvector1))
  {
    SUMvector[n]=TPvector1[n]*log10(TPvector1[n]/(TPvectorX[unlist(strsplit(names(TPvector1)[n],"i"))[1]]
*TPvectorY[unlist(strsplit(names(TPvector1)[n],"i"))[length(strsplit(names(TPvector1)[1],"i")[[1]])]]))
  }
  return(sum(SUMvector))
}

> mi(X,Y)
[1] 0.0005991637
```

Written in R (http://www.r-project.org/).

Mutual information

- Value (6x10ˆ-4) may be close to the 'actual' MI of the system but it may be, as well, too small or too high.
- However, ”a time-series lenght of only ten symbols provides a relatively close approximation of the population mutual information for the system, with an average deviation of around 0.04.” [1]

Mutual information

- Shortcomings of MI:
  - MI is not effective at predicting future events from current data: it is symmetric, M(X,Y)=M(Y,X).
  - It does not indicate which way the information is flowing.
- These shortcomings may be remedied by *time shifting* one of the variables.
  - *Transfer Entropy* (TE) (Schreiber 2000 [2]) is based on rates of entropy change, it captures some of the dynamics of a system.

Transfer entropy

- Suppose two systems which generates events.
- We define an entropy rate which is the amount of additional information required to represent the value of the next observation of one of the systems:

$$h_1 = -\sum_{x_{n+1}} p(x_{n+1}, x_n, y_n) \log_a p(x_{n+1} | x_n, y_n)$$

- Suppose that value of observation x_{n+1} was not dependent on the current observation y_n :

$$h_2 = -\sum_{x_{n+1}} p(x_{n+1}, x_n, y_n) \log_a p(x_{n+1} | x_n)$$

Transfer entropy

- Now, the quantity h_1 represents the entropy rate for the two systems, and h_2 represents the entropy rate assuming that x_{n+1} is independent of y_n.Thus, we get *transfer entropy*:

$$
\begin{aligned}
h_2 - h_1 &= -\sum_{x_{n+1},x_n,y_n} p(x_{n+1},x_n,y_n) \log_a p(x_{n+1}|x_n) \\
&\quad + \sum_{x_{n+1},x_n,y_n} p(x_{n+1},x_n,y_n) \log_a p(x_{n+1}|x_n,y_n) \\
&= \sum_{x_{n+1},x_n,y_n} p(x_{n+1},x_n,y_n) \log_a \left(\frac{p(x_{n+1}|x_n,y_n)}{p(x_{n+1}|x_n)} \right)
\end{aligned}
$$

Transfer entropy

- There are actually two equations for the transfer entropy, because it has an inherent asymmetry in it.

$$
\begin{aligned}
T_{J \to I} &= \sum_{x_{n+1}, x_n, y_n} p(x_{n+1}, x_n, y_n) \log \left(\frac{p(x_{n+1} | x_n, y_n)}{p(x_{n+1} | x_n)} \right) \\
T_{I \to J} &= \sum_{y_{n+1}, x_n, y_n} p(y_{n+1}, x_n, y_n) \log \left(\frac{p(y_{n+1} | x_n, y_n)}{p(y_{n+1} | y_n)} \right)
\end{aligned}
$$

Transfer entropy

- With substitutions

$$p(x_{n+1}|x_n, y_n) = p(x_{n+1}, x_n, y_n)/p(x_n, y_n)$$

$$p(x_{n+1}|x_n) = p(x_{n+1}, x_n)/p(x_n)$$

our equations become

$$T_{J\to I} = \sum_{x_{n+1},x_n,y_n} p(x_{n+1}, x_n, y_n) \log\left(\frac{p(x_{n+1}, x_n, y_n)\cdot p(x_n)}{p(x_n, y_n)\cdot p(x_{n+1}, x_n)}\right)$$

$$T_{I\to J} = \sum_{y_{n+1},x_n,y_n} p(y_{n+1}, x_n, y_n) \log\left(\frac{p(y_{n+1}, x_n, y_n)\cdot p(y_n)}{p(x_n, y_n)\cdot p(y_{n+1}, y_n)}\right)$$

Transfer entropy

- Let's use our previous data as an example:

  1: 1110110100110011101101

  2: 1010101011011011011001

- First determine $p(x_{n+1}, x_n, y_n)$

$$p(0_x, 0_x, 0_y) = 0.0$$

$$p(0_x, 0_x, 1_y) = 0.0952381$$

$$p(0_x, 1_x, 0_y) = 0.190476$$

$$p(0_x, 1_x, 1_y) = 0.0952381$$

$$p(1_x, 0_x, 0_y) = 0.142857$$

$$p(1_x, 0_x, 1_y) = 0.142857$$

$$p(1_x, 1_x, 0_y) = 0.0952381$$

$$p(1_x, 1_x, 1_y) = 0.238095$$

Transfer entropy

- Then we calculate $p(x_{n+1}, x_n)$ and $p(x_n, y_n)$

$$\begin{aligned} p(0,0) &= 0.0952381 \\ p(0,1) &= 0.285714 \\ p(1,0) &= 0.285714 \\ p(1,1) &= 0.333333 \end{aligned}$$

$$p(0,0) = 0.136364$$

$$p(0,1) = 0.227273$$

$$p(1,0) = 0.272727$$

$$p(1,1) = 0.363636$$

and finally $p(x)$

$$p(0) = 0.363636$$

$$p(1) = 0.636364$$

```
trent<-function(Y,X,s=1){
    L4=L1=length(X)-s # Lengths of vectors.
    L3=L2=length(X)
    #--------------------#
    # 1. p(Xn+s,Xn,Yn): #
    #--------------------#
    TPvector1=rep(0,L1) # Init.
    for(i in 1:L1)
    {
            TPvector1[i]=paste(c(X[i+s],"i",X[i],"i",Y[i]),collapse="") # "addresses"
    }
    TPvector1T=table(TPvector1)/length(TPvector1) # Table of probabilities.
    #-----------#
    # 2. p(Xn): #
    #-----------#
    TPvector2=X
    TPvector2T=table(X)/sum(table(X))
    #--------------#
    # 3. p(Xn,Yn): #
    #--------------#
    TPvector3=rep(0,L3)
    for(i in 1:L3)
    {
            TPvector3[i]=paste(c(X[i],"i",Y[i]),collapse="") # addresses
    }
    TPvector3T=table(TPvector3)/length(TPvector2)
    #----------------#
    # 4. p(Xn+s,Xn): #
    #----------------#
    TPvector4=rep(0,L4)
    for(i in 1:L4)
    {
            TPvector4[i]=paste(c(X[i+s],"i",X[i]),collapse="") # addresses
    }
    TPvector4T=table(TPvector4)/length(TPvector4)
    #--------------------------#
    # Transfer entropy T(Y->X) #
    #--------------------------#
        SUMvector=rep(0,length(TPvector1T))
    for(n in 1:length(TPvector1T))
    {
        SUMvector[n]=TPvector1T[n]*log10((TPvector1T[n]*TPvector2T[(unlist(strsplit(names(TPvector1T)[n],"i")))[2]])/(TPvector3T[paste
((unlist(strsplit(names(TPvector1T)[n],"i")))[2],"i",(unlist(strsplit(names(TPvector1T)[n],"i")))[3],sep="",collapse="")]*TPvector4T
[paste((unlist(strsplit(names(TPvector1T)[n],"i")))[1],"i",(unlist(strsplit(names(TPvector1T)[n],"i")))[2],sep="",collapse="")]))
    }
    return(sum(SUMvector))
} # End of the trent-function.
> trent(X,Y); trent(Y,X)
[1] 0.01131521
[1] 0.0440033
```

Transfer entropy

- >X=as.character(c(1,1,1,0,1,1,0,1,0,0,1,1,0,0,1,1,1,0,1,1,0,1))
- >Y=as.character(c(1,0,1,0,1,0,1,0,1,1,0,1,1,0,1,1,0,1,1,0,0,1))
- > trent(X,Y)
- [1] 0.01131521
- > trent(Y,X)
- [1] 0.0440033
- Thus, system Y adds 0.044 digits of predictability to system X, and system X adds 0.011 digits of predicability to Y.

Transfer entropy

- ”TE is more adequate (than MI or time-delayed MI) for determining the direction of inf. flow between two coupled processes.” [3]
- ”By means of any directional measure of interdependence within bivariate signals, one cannot prove the presence of actual coupling strengths, nor exclude the influences of many other systems.” (Feldmann & Bhattachary 2004)

Sample applications

- Causal relations between pairs of genes. (Tung et al 2007)
  - Many of our findings are supported by biological evidences.
- Dependency between heart and breath rate. [4]
  - Heart rate influences the breath rate rather than vice versa.
- Information Transfer Between Auditory Cortical Neurons. (Gourevitch & Eggermont 2006)
  - In conclusion, normalized transfer entropy or NTE has promising features that should make it useful for neural networks analysis.

Sample applications

- The strength and the direction of information transfer in the US stock market. (Baek et al 2005)
  - Our entropy analysis shows that the companies related with energy industries such as oil, gas, and electricity influence the whole market.
- The causality inherent in the face-to-face interaction, detecting the causality between perception and action variables in human-robot interaction. (Sumioka, Yoshikava 2007)
  - Transfer entropy helped a robot to detect important variables that constitute a causal structure inherent in the interaction.

Sample applications

- Calcium signaling under different cellular conditions. (Pahle et al. 2008)
  - Even though the estimation of transfer entropy from time series is tricky and there are still some unsolved issues, it is a promising tool not only for the quantification of information transfer in biochemical networks, but also, for instance, to distinguish between different stochastic time series where a pure visual investigation is difficult.
- TE in ecological monitoring programs. (L.J. Moniz et al. 2007)

TE applied to music analysis

- Algorithms and math can be used to analyse and compose music.
- Musical information can be converted to discrete time-series.
- Kulp & Schlingmann [5] estimated the rate of information transfer between string instruments of Beethoven symphonies.
  - They changed note pitches and durations to time-series:  = $\{440 \quad 440 \quad 523.35 \quad 523.35 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0\}$

TE applied to music analysis

(Kulp & Schlingmann 2007)

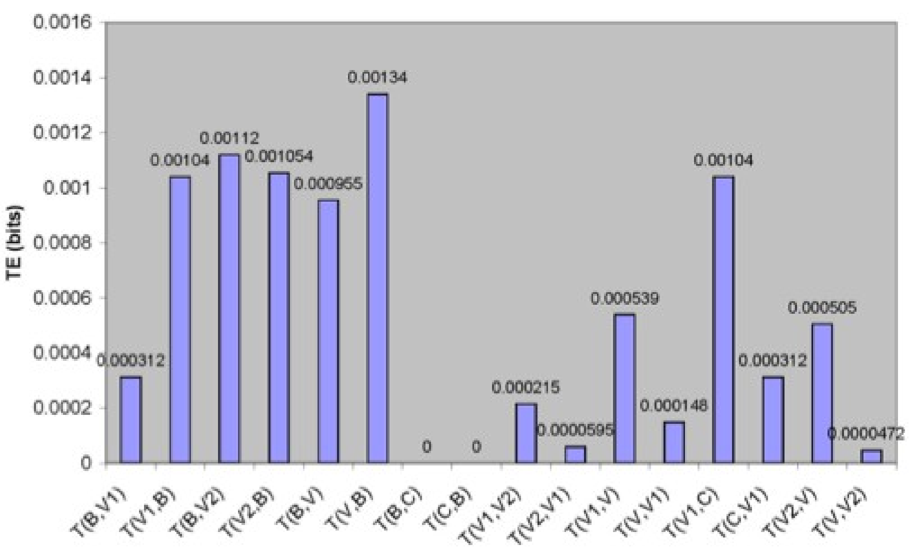


Figure 3: The transfer entropy analysis of the string section of Beethoven's First Symphony. Note that: V1 = first violin, V2 = second violin, V = viola, C = cello, and B = bass.

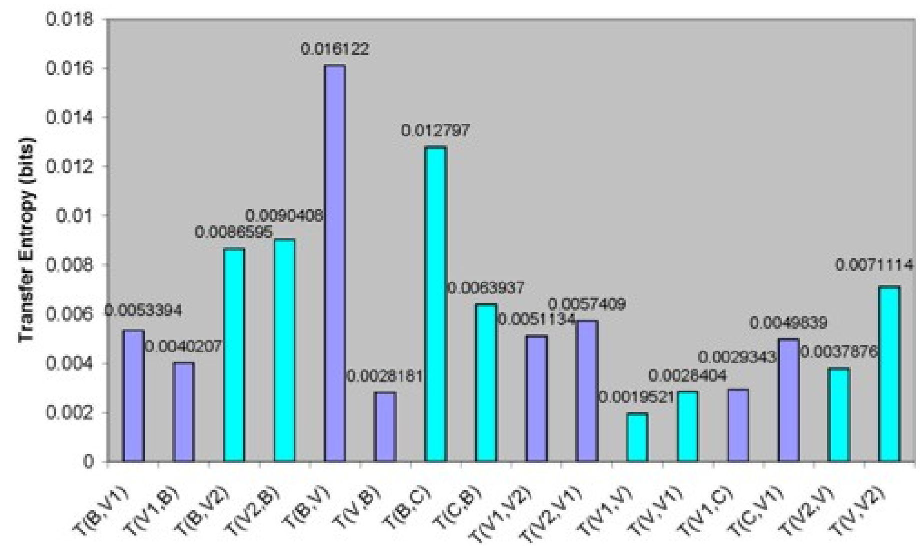


Figure 4: The transfer entropy analysis of the string section of Beethoven's Fifth Symphony.

TE applied to music analysis

- TE was now applied to J.S. Bach's (1685-1750) contrapuntally arranged themes.
  - Bach wrote his themes (melodies) in a way that their positions can be changed, i.e. every part can operate, for example, as bass. This is called *multiple counterpoint*. (Thus, the number of contrapuntal combinations is n!)
  - Are they equal, or is some of them in a more determining position?
  - Could we assume that their composition order defines their musically 'hierarchical' position?

TE applied to music analysis

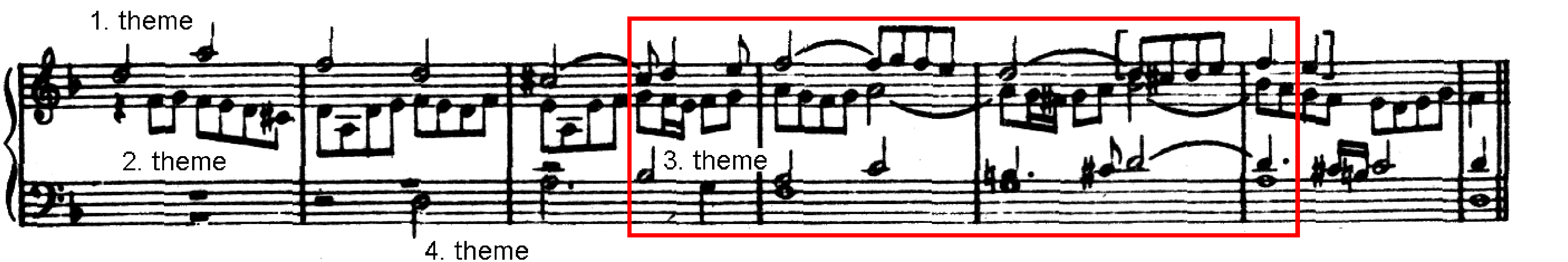
J.S. Bach, Kunst der Fuge. Themes from the last, unfinished fugue.

TE applied to music analysis

- The last fugue of Kunst der Fuge remained unfinished.
- KdF is based on the first theme (see the previous slide) which does not, in fact, originally appear in the last fugue. The quadruple counterpoint, seen in the previous slide, was discovered by Gustav Nottebohm in 1880.
- The original manuscript ends soon after the third subject (B-A-C-H) has been introduced.

TE applied to music analysis

- 4 themes were converted to MIDI-pitches:




```
T1=c(73,73,74,74,74,74,76,76,77,77,77,77,77,77,77,77,77,77,79,79,77,77,76,76,74,
74,74,74,74,74,74,74,74,74,73,73,74,74,76,76,77,77,77,77)
T2=c(67,67,65,64,65,65,67,67,69,69,67,67,65,65,67,57,69,69,69,69,69,69,69,69,69,
69,67,66,67,67,69,69,70,70,70,70,70,70,70,70,70,70,69,69)
T3=c(58,58,58,58,58,58,58,58,57,57,57,57,57,57,57,57,60,60,60,60,60,60,60,60,59,
59,59,59,59,59,61,61,62,62,62,62,62,62,62,62,62,62,62,62)
T4=c(57,57,57,57,55,55,55,55,53,53,53,53,53,53,53,53,53,53,53,53,53,53,53,53,55,
55,55,55,55,55,55,55,55,55,55,55,55,55,55,55,57,57,57,57)
```

TE applied to music analysis

- TE:s between the themes were calculated using time shifts (n+1)...(n+8). Then rowSums of $T_{I\text{->}J}$ were calculated:
  - 3rd theme attains in every case the greatest value.

Time shift	T1	T2	T3	T4
n+1	0.214	0.182	**0.227**	0.154
n+2	0.521	0.391	**0.530**	0.319
n+3	0.427	0.378	**0.530**	0.328
n+4	0.540	0.551	**0.772**	0.483
n+5	0.448	0.523	**0.728**	0.342
n+6	0.644	0.732	**0.905**	0.424
n+7	0.511	0.603	**0.868**	0.362
n+8	0.619	0.648	**1.009**	0.403

H(T1)= 1.97
H(T2)= 2.23
H(T3)= 2.45
H(T4)= 1.49

$T_{I\text{->}J} - T_{J\text{->}I}$, time shift=n+1:

T1	T2	T3	T4
0.027	-0.138	0.135	-0.024

Questions

- What does TE actually tell about the music in such an analysis example as presented before? (OK, first, the sequence is surely too short...)
- Which of the musical parameters would serve best for the purpose: actual pitches, intervals between the notes or rhythms?
- Could we use multidimensional features as well, by combining information of all parameters mentioned above?
- How about using higher (than first) order chains?

Questions

- What happens if we move a sliding window (using overlapping windows) throughout a whole piece and calculate TE in each moment?

- What if all theme occurrences are checked against other counterthemes and TE-means are calculated for each theme?

Proposals for analysis objects

- TE between two improvising (and communicating) drummers.
- TE between improvising jazz musicians.
- TE between two singing (and communicating) birds?
- ...

References

[1] Davis, A. 2002. *Small Sample Effects on Information-Theoretic Estimates*. Thesis, College of William and Mary in Virginia.

[2] Schreiber, T. 2000. *Measuring Information Transfer*. Physical Review Letters 85: 461-464.

[3] Kaiser A. & Schreiber T. 2002. *Information transfer in continuous processes*. Physica D 166: 43–62.

[4] Bauer M. et al. 2004. *Specifying the directionality of fault propagation paths using transfer entropy*. DYCOPS7, Boston.

[5] Kulp C.W. & Schlingmann D. 2007. *Composition and Analysis of Music Using Mathematica*. MCM2007, Berlin.

THANKS!