HOMOGENEITY ANALYSIS OF CATEGORICAL TIME SERIES

JAN DE LEEUW

Abstract. Meet the abstract. This is the abstract.

1. Data

Suppose we have a *time series* of the form

A A G C A T T T A A C G T ⋯

If we choose a window width $m = 4$, and ignore whatever comes after the first 13 elements of the series, we can expand the series to the *Hankel matrix*

A	A	G	C
A	G	C	A
G	C	A	T
C	A	T	T
A	T	T	T
T	T	T	A
T	T	A	A
T	A	A	C
A	A	C	G
A	C	G	T

The Hankel matrix can be expanded to the *indicator matrix G*

Date: Wednesday 10th February, 2010 — 16h 54min — Typeset in Kepler.

Key words and phrases. Template, LaTeX.

A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T
1	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0
1	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0
0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	1
0	1	0	0	1	0	0	0	0	0	0	1	0	0	0	1
1	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1
0	0	0	1	0	0	0	1	0	0	0	1	1	0	0	0
0	0	0	1	0	0	0	1	1	0	0	0	1	0	0	0
0	0	0	1	1	0	0	0	1	0	0	0	0	1	0	0
1	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0
1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1

The crossproduct of the indicator matrix is the *Burt matrix C*, which has the *marginals D* as its diagonal.

	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T
A	5	0	0	0	2	1	1	1	0	2	2	1	1	1	1	2
C	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0	1
G	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	1
T	0	0	0	3	1	0	0	2	2	0	0	1	2	1	0	0
A	2	1	0	1	4	0	0	0	1	1	1	1	0	2	1	1
C	1	0	1	0	0	2	0	0	1	0	1	0	0	0	0	2
G	1	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0
T	1	0	0	2	0	0	0	3	1	0	0	2	2	0	0	1
A	0	0	1	2	1	1	0	1	3	0	0	0	1	1	0	1
C	2	0	0	0	1	0	1	0	0	2	0	0	1	0	1	0
G	2	0	0	0	1	1	0	0	0	0	2	0	0	1	0	1
T	1	1	0	1	1	0	0	2	0	0	0	3	1	0	0	2
A	1	0	0	2	0	0	1	2	1	1	0	1	3	0	0	0
C	1	0	0	1	2	0	0	0	1	0	1	0	0	2	0	0
G	1	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0
T	2	1	1	0	1	2	0	1	1	0	1	2	0	0	0	4

In homogeneity analysis we solve the eigenproblem $Cy = m\lambda^2 Dy$, which is the same thing as solving the singular value problem $Gy = m\lambda x$ and $G'x = \lambda Dy$. This is probably identical to the way in which SSA handles multidimensional numerical series.

The Burt matrix we have calculated is close to a *block Toeplitz matrix*, in fact it is a block Toeplitz matrix except for boundary effects. The block Toeplitz matrix T can be calculated from the marginals of the series.

	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T
A	5	0	0	0	2	1	1	1	0	2	2	1	1	1	1	2
C	0	2	0	0	1	0	1	0	0	0	0	2	0	0	0	1
G	0	0	2	0	0	1	0	1	1	0	0	0	0	0	0	1
T	0	0	0	4	1	0	0	2	2	0	0	1	2	1	0	0
A	2	1	0	1	5	0	0	0	2	1	1	1	0	2	2	1
C	1	0	1	0	0	2	0	0	1	0	1	0	0	0	0	2
G	1	1	0	0	0	0	2	0	0	1	0	1	1	0	0	0
T	1	0	1	2	0	0	0	4	1	0	0	2	2	0	0	1
A	0	0	1	2	2	1	0	1	5	0	0	0	2	1	1	1
C	2	0	0	0	1	0	1	0	0	2	0	0	1	0	1	0
G	2	0	0	0	1	1	0	0	0	0	2	0	0	1	0	1
T	1	2	0	1	1	0	1	2	0	0	0	4	1	0	0	2
A	1	0	0	2	0	0	1	2	2	1	0	1	5	0	0	0
C	1	0	0	1	2	0	0	0	1	0	1	0	0	2	0	0
G	1	0	0	0	2	0	0	0	1	1	0	0	0	0	2	0
T	2	1	1	0	1	2	0	1	1	0	1	2	0	0	0	4

Both have some advantages and some drawbacks. The Burt table has the property that the rows and columns of the block add up to the diagonal elements, but the matrix is not exactly Block Toeplitz. In the Block Toeplitz matrix the rows and columns of the blocks do not add up to the diagonal elements. The differences between the matrices are minimal, expecially for long series with relatively small

windows. Again, for the Block Toeplitz matrix we can solve the eigen problem $Ty = mSy$, with $S = \mathbf{diag}(T)$.

2. Markov

If a series is exactly Markovian, then the block matrix (with window width four) is

$$\begin{bmatrix} D_1 & D_1P & D_1P^2 & D_1P^3 \\ P'D_1 & D_2 & D_2P & D_2P^2 \\ P'^2D_1 & D_2P & D_3 & D_3P \\ P'^3D_1 & D_2P'^2 & D_3P' & D_4 \end{bmatrix}$$

If the chain is stationary, all the D_j are equal to, say, D_∞. If the chain is time-reversible, then $PD_\infty = D_\infty P$. Under those conditions homogeneity analysis simply computes the spectral decomposition of

$$\begin{bmatrix} I & P & P^2 & P^3 \\ P' & I & P & P^2 \\ P'^2 & P' & I & P \\ P'^3 & P'^2 & P' & I \end{bmatrix}$$

If in addition the chain is symmetric, i.e. if $P = P'$, then

$$\begin{bmatrix} I & P & P^2 & P^3 \\ P & I & P & P^2 \\ P^2 & P & I & P \\ P^3 & P^2 & P & I \end{bmatrix}$$

In this case the homogeneity analysis is simply the eigen analysis of the symmetric matrix P, presented in a more complicated form. In fact, the eigenvalues are the eigenvalues of the direct sum of four

(number of states, order of P) Toeplitz matrices of order four (window width) that have the form

$$\begin{bmatrix} 1 & \lambda & \lambda^2 & \lambda^3 \\ \lambda & 1 & \lambda & \lambda^2 \\ \lambda^2 & \lambda & 1 & \lambda \\ \lambda^3 & \lambda^2 & \lambda & 1 \end{bmatrix}$$

where λ is one of the four eigenvalues for P. The eigenvectors will have the familiar interlocked sine-cosine pattern.

Many of the results in this section remain true for a much more general process $D_i^{-\frac{1}{2}} C_{ij} D_j^{-\frac{1}{2}} = K_i \Lambda_{ij} K_j'$ with $K_i' K_i = I$ for all i, and with all Λ_{ij} diagonal. Since $C_{ij} = D_i P_{ij}$ this implies

$$P_{ij} P_{j\ell} = D_i^{-\frac{1}{2}} K_i \Lambda_{ij} \Lambda_{j\ell} K_\ell' D_\ell^{\frac{1}{2}},$$

while

$$P_{i\ell} = D_i^{-1} C_{i\ell} = D_i^{-\frac{1}{2}} K_i \Lambda_{i\ell} K_\ell' D_\ell^{\frac{1}{2}}.$$

Thus in this case $\Lambda_{ij} \Lambda_{j\ell} = \Lambda_{i\ell}$ guarantees the Markovian semi-group property.

3. Multinormal

This has been studied in detail. If the process is multinormal and mean/variance stationary (not necessarily covariance stationary), then the C_{ij} will be bivariate normal, and they can be diagonalized by the Hermite-Chebyshev polynomials (this assumes a continuum of states, which will have to be discreticized for computation). The previous results apply with Λ_{ij} a diagonal matrix with the powers of ρ_{ij},

the autocorrelation between times i and j. For Markovity it is now necessary and sufficient that $\rho_{ij}\rho_{j\ell} = \rho_{i\ell}$.

Department of Statistics, University of California, Los Angeles, CA 90095-1554

E-mail address, Jan de Leeuw: deleeuw@stat.ucla.edu

URL, Jan de Leeuw: http://gifi.stat.ucla.edu