Measuring Serial Dependence in Categorical Time Series

Christian H. Weiß

University of Würzburg

Institute of Mathematics

Department of Statistics

This talk is based on the paper

Weiß, C.H., Göb, R.:

Measuring Serial Dependence in Categorical Time Series.

Preprint 265, University of Würzburg, 2006.

All references mentioned in this talk correspond to the references in this paper.

Measuring Serial Dependence

Motivation

An intrinsic feature of a time series is that, typically, adjacent observations are dependent. The nature of this dependence among observations of a time series is of considerable practical interest. Time series analysis is concerned with techniques for the analysis of this dependence. (Box et al., 1994, p. 1)

Cardinal time series:

Convenient measure of serial dependence:

(Partial) Autocorrelation.

Categorical time series:

Measures of serial dependence?

Concepts of Stationarity:

- Strict stationarity: Applicable to any type of time series. But too strict for practice.

- Cardinal time series: Weak stationarity coordinated with autocorrelation.

- Categorical time series: Weak stationarity?

Measuring Dispersion of Categorical Random Variables

A Review

Intuitive understanding of dispersion:

X shows large dispersion

$\approx$

High uncertainty about the outcome of X

$\Rightarrow$ Uncertainty of Categorical Random Variable?

Two extreme cases:

Uniform distribution:

Maximal uncertainty about the outcome of X.

One-point distribution:

Perfect certainty about the outcome of X.

$\Rightarrow$ Hallmarks for definition of any measure of dispersion!

Numerous contributions in literature:

- Desirable properties, suggestions on measures:
  Uschner (1987), Vogel & Kiesl (1999), and many more.

- Dispersion in the discrete ordinal case: Kiesl (2003).

Common standardized measures of dispersion:

Gini index: $\nu_G(X) \ := \ \frac{m}{m-1}\,\big(1 \ - \ \sum_{j=1}^{m} p_j^2\big).$

Entropy: $\nu_E(X) \ := \ -\frac{1}{\ln m}\ \sum_{j=1}^{m} p_j \ln p_j.$

Chebycheff dispersion: $\nu_C(X) \ := \ \frac{m}{m-1}\,\big(1 \ - \ \max_j p_j\big).$

Important properties of these measures:

- continuous and symmetric functions of distribution $p_i = \mathsf{P}(X = x_i)$,
- range $[0; 1]$,
- maximum value 1 in case of uniform distribution,
- minimal value 0 in case of one-point distribution,
- inequality: $\frac{m}{m-1}\,(1 \;-\; \min_j p_j) \;\geq\; \nu_G(X) \;\geq\; \nu_C(X)$.

Measuring Dependence of Categorical Random Variables

Desirable Properties

Some notational remarks:

- X, Y categorical random variables with range $\mathcal{V}_x = \{x_1, \ldots, x_{m_x}\}$ resp. $\mathcal{V}_y = \{y_1, \ldots, y_{m_y}\}$.
- marginal distributions: $\mathsf{P}(X = x_i) = p_{x,i}$, $\mathsf{P}(Y = y_j) = p_{y,j}$.
- Ranges chosen such that $p_{x,i}, p_{y,j} > 0$.
- $p_{ij} = \mathsf{P}(X = x_i, Y = y_j)$ joint probability, $p_{i|j} = \mathsf{P}(X = x_i | Y = y_j) = \frac{p_{ij}}{p_{y,j}}$ conditional probability.

Extreme cases:

- X, Y *(stochastically) independent* iff $p_{ij} = p_{x,i} \cdot p_{y,j}$.

- X *perfectly depends* on Y iff for every $j = 1, \ldots, m_y$: conditional distribution of X, conditioned on $Y = y_j$, is a one-point distribution.

Properties of Perfect Dependence:

- If X depends perfectly on Y, then $m_x \leq m_y$.
- Let $m_x = m_y$. If X depends perfectly on Y, then Y depends perfectly on X.

Perfect dependence is in general a *non-symmetric* relation.

Essential properties of measure $A(X,Y)$ of dependence:

- $A(X,Y)$ only depends on m_x, m_y and $p_{x,i}, p_{y,j}, p_{ij}$, continuous function thereof.
- X, Y are independent $\begin{matrix}\Rightarrow \\ \Leftarrow\end{matrix}$ $A(X,Y) = 0$.
- Fixed m_x, m_y and $p_{x,i}, p_{y,j}$: $A(X,Y)$ has range $[0; a]$.
- X depends perfectly on Y $\begin{matrix}\Rightarrow \\ \Leftarrow\end{matrix}$ $A(X,Y) = a$.
- The measure is symmetric in X and Y.

Measuring Dependence of Categorical Random Variables

Proportional Reduction of Variation

Concept: $A_\nu(X|Y) \;:=\; \frac{\nu(X)-E[\nu(X|Y)]}{\nu(X)} \;=\; 1-\frac{E[\nu(X|Y)]}{\nu(X)}.$

- *Goodman and Kruskal's* τ based on the *Gini index*:

$$A_\nu^{(\tau)}(X|Y) \;=\; \frac{\sum_{i=1}^{m_x}\sum_{j=1}^{m_y}\frac{(p_{ij}\;-\;p_{x,i}\;p_{y,j})^2}{p_{y,j}}}{1-\sum_{i=1}^{m_x}p_{x,i}^2}.$$

- The *uncertainty coefficient* based on *entropy*:

$$A_\nu^{(u)}(X|Y) \;=\; -\frac{\sum_{i=1}^{m_x}\sum_{j=1}^{m_y}p_{ij}\ln\left(\frac{p_{ij}}{p_{x,i}p_{y,j}}\right)}{\sum_{i=1}^{m_x}p_{x,i}\ln p_{x,i}}.$$

Properties:

- $A(X, Y)$ only depends on m_x, m_y and $p_{x,i}, p_{y,j}, p_{ij}$, continuous function thereof.
- X, Y are independent $\quad\Leftrightarrow\quad A(X, Y) = 0$.
- Fixed m_x, m_y and $p_{x,i}, p_{y,j}$: $A(X, Y)$ has range $[0; 1]$.
- X depends perfectly on $Y \quad\Leftrightarrow\quad A(X, Y) = 1$.

- The measure is *not* symmetric in X and Y.

- *Goodman and Kruskal's* λ based on *Chebycheff disp.*:

$$A_\nu^{(\lambda)}(X|Y) \;=\; \frac{\sum_{j=1}^{m_y} \max_i p_{ij} - \max_i p_{x,i}}{1 - \max_i p_{x,i}}.$$

Properties: Like above, but:

- X, Y are independent $\quad\Rightarrow\quad A(X,Y) = 0$.

  The inverse is not true!

Measuring Dependence of Categorical Random Variables

Sample Statistics

Popular examples:

- *Pearsons's χ^2-statistic*: $m := \min(m_x, m_y)$.

$$\mathsf{X}_n^2(X,Y) = n \sum_{i=1}^{m_x} \sum_{j=1}^{m_y} \frac{(p_{ij} - p_{x,i}p_{y,j})^2}{p_{x,i}p_{y,j}}.$$

- Φ^2 *measure*: $\Phi^2(X,Y) = \frac{\mathsf{X}_n^2(X,Y)}{n}$.
- *Sakoda's measure*: $p^*(X,Y) = \sqrt{\frac{m}{m-1} \cdot \left(1 - \frac{1}{1+\Phi^2(X,Y)}\right)}$.
- *Cramér's v*: $v(X,Y) := \frac{\Phi(X,Y)}{\sqrt{m-1}}$.

Properties:

- $A(X,Y)$ only depends on m_x, m_y and $p_{x,i}, p_{y,j}, p_{ij}$, continuous function thereof.
- X, Y are independent $\quad\Leftrightarrow\quad A(X,Y) = 0$.
- Fixed m_x, m_y and $p_{x,i}, p_{y,j}$: $A(X,Y)$ has range $[0; a]$.
- X depends perfectly on $Y \quad\Leftrightarrow\quad A(X,Y) = a$.
- The measure is symmetric in X and Y.

Sakoda's measure, Cramér's v: $a = 1$.

Measuring Dependence of Categorical Random Variables

Signed Dependence

Motivation:

Cardinal case: Positive and negative correlation.

Definition:

- X, Y with identical range $\{z_1, \ldots, z_m\}$.
- X, Y *perfectly positively dependent*,
  if they are perfectly dependent and
  if $p_{i|i} = 1$ for all i.
- X, Y *perfectly negatively dependent*
  if perf. dep. and $p_{i|i} = 0$ for all i.

Essential properties of measure $A(X,Y)$ of dependence:

- $A(X,Y)$ only depends on m_z and $p_{x,i}, p_{y,j}, p_{ij}$, continuous function thereof.
- X, Y are independent $\underset{\Leftarrow}{\Rightarrow}$ $A(X,Y) = 0$.
- Fixed m_z, $p_{x,i}, p_{y,j}$: $A(X,Y)$ has range $[l; u]$, $l < 0 < u$.
- X, Y *perfectly positively dependent* $\underset{\Leftarrow}{\Rightarrow}$ $A(X,Y) = u$.
- X, Y *perfectly negatively dependent* $\underset{\Leftarrow}{\Rightarrow}$ $A(X,Y) = l$.
- The measure is symmetric in X and Y.

Cohen's κ: $\kappa(X,Y) \;:=\; \frac{\sum_{j=1}^{m}(p_{jj}-p_{x,j}p_{y,j})}{1-\sum_{j=1}^{m}p_{x,j}p_{y,j}}$.

Properties:

- $\kappa(X,Y)$ only depends on m_z and $p_{x,i}, p_{y,j}, p_{ij}$, continuous function thereof.
- X, Y are independent $\quad\Rightarrow\quad \kappa(X,Y)=0$.
- Fixed m_z, $p_{x,i}, p_{y,j}$: κ has range $[-\frac{\sum_{j=1}^{m}p_{x,j}p_{y,j}}{1-\sum_{j=1}^{m}p_{x,j}p_{y,j}}\;;\;1]$.
- X, Y *perfectly positively dependent* $\quad\Leftrightarrow\quad \kappa(X,Y)=1$.
- X, Y *perfectly negatively dependent* $\quad\Rightarrow\quad \kappa$ minimal.
- The measure is symmetric in X and Y.

Serial Dependence in Categorical Time Series

Weak Stationarity

- Previously defined measures all applicable to categorical time series: $A(X_t, X_{t-k})$

- Important simplification: X_t and X_{t-k} have same range
  $\Rightarrow$ Perfect dependence is symmetric relation,
  $\Rightarrow$ Signed perfect dependence defined.

- In general: $A(X_t, X_{t-k})$ depends on t.

Weak forms of stationarity for categorical processes:

- Marginal stationarity.
- Harris & McGee (2004): affects marginal distribution.
- *Measure A stationarity*: $A(X_t, X_{t-k})$ invariant in t.
  But no standard measure exists.
- *Bivariate stationarity*: Joint distribution of X_{t-k}, X_t invariant in t.
  $\Rightarrow A(X_t, X_{t-k})$ invariant in t for any A, i. e.,
  'Autodependence': $A(k) \;:=\; A(X_t, X_{t-k})$.

Bivariate stationarity $\Rightarrow$ Simplifications:

- Goodman's τ: $A_\nu^{(\tau)}(k) \;=\; \frac{\sum_{i,j=1}^m \frac{p_{ij}(k)^2}{p_j} - \sum_{i=1}^m p_i^2}{1-\sum_{i=1}^m p_i^2}$.

- Pearson's χ^2-statistic: $\mathrm{X}_n^2(k) \;=\; n\sum_{i,j=1}^m \frac{(p_{ij}(k)-p_ip_j)^2}{p_ip_j}$.

- Cramér's v: $v(k) \;=\; \frac{\mathrm{X}_n^2(k)}{\sqrt{n(m-1)}}$.

- Cohen's κ: $\kappa(k) \;=\; \frac{\sum_{j=1}^m (p_{jj}(k)-p_j^2)}{1-\sum_{j=1}^m p_j^2}$.

Serial Dependence in Categorical Time Series

An Example

NDARMA model of Jacobs & Lewis (1983):

- $(\varepsilon_t)_{\mathbb{Z}}$ i.i.d. with marginal by $\mathrm{P}(\varepsilon_t = x_j) = \pi_j$.
- i.i.d. decision variables $\boldsymbol{D}_t = (\alpha_{1,t}, \ldots, \alpha_{p,t}, \beta_{0,t}, \ldots, \beta_{q,t}) \sim MULT(1; \phi_1, \ldots, \phi_p, \varphi_0, \ldots, \varphi_q)$.
- $X_t \;=\; \alpha_{1,t}\, X_{t-1} + \ldots + \alpha_{p,t}\, X_{t-p} + \beta_{0,t}\, \varepsilon_t + \ldots + \beta_{q,t}\, \varepsilon_{t-q}$.

- $\mathrm{P}(X_{t_1} = x_{i_1}, X_{t_2} = x_{i_2}) \;=$
  $\pi_{i_1}\pi_{i_2}\,\big(1 - \mathrm{Corr}[X_{t_1}, X_{t_2}]\big) \;+\; \delta_{i_1 i_2}\,\pi_{i_1}\,\mathrm{Corr}[X_{t_1}, X_{t_2}]$
  $\Rightarrow$ Positive dependence.

Corr$[X_{t_1}, X_{t_2}]$ always interpretable, and:

$$\begin{aligned} \text{Corr}[X_{t_1}, X_{t_2}] \quad &= \kappa(X_{t_1}, X_{t_2}) \\ &= \upsilon(X_{t_1}, X_{t_2}) \\ &= \sqrt{A_\nu^{(\tau)}(X_{t_1}, X_{t_2})}. \end{aligned}$$

Under bivariate stationary:

Estimation of $\kappa,\ \upsilon,\ A_\nu^{(\tau)}$ possible

$\Rightarrow$ Check for model adequacy.

$(X_t)_{\mathbb{Z}}$ bivariate stationary NDARMA(p, q) process,

with 'autocorrelation' $\rho(k) \; = \;$ Corr$[X_t, X_{t-k}]$.

Yule-Walker equations:

$$\rho(k) \; = \; \sum_{j=1}^{p} \phi_j \; \rho(|k-j|) \; + \; \sum_{i=1}^{q-k} \varphi_{i+k} \; r(i),$$

$\Rightarrow$ Model estimation.

Also partial autocorrelation for identifying DAR(p) model.

Bovine leukemia virus:

Lag k	$\widehat{\kappa}(k)$	$\widehat{\upsilon}(k)$	$\sqrt{\widehat{A}_{\nu}^{(\tau)}(k)}$	$\widehat{\rho}_{\mathsf{p}}(k)$
1	0.0804	0.1134	0.1118	0.0804
2	0.0248	0.0445	0.0447	0.0185
3	0.0008	0.0281	0.0299	−0.0026
4	−0.0065	0.0222	0.0232	−0.0069
5	−0.0151	0.0294	0.0300	−0.0141

$\widehat{\pi}_a = 0.220,\ \widehat{\pi}_c = 0.331,\ \widehat{\pi}_g = 0.210, \widehat{\pi}_t = 0.239.$

DAR(2): $\widehat{\varphi}_0 = 0.903,\ \widehat{\phi}_1 = 0.079,\ \widehat{\phi}_2 = 0.019.$

Thank You

for Your Interest!

Christian H. Weiß

University of Würzburg

Institute of Mathematics

Department of Statistics