Renewal Methods of Generating Stationary Count Time Series

Robert Lund

Department of Mathematical Sciences
Clemson University
Lund@Clemson.edu

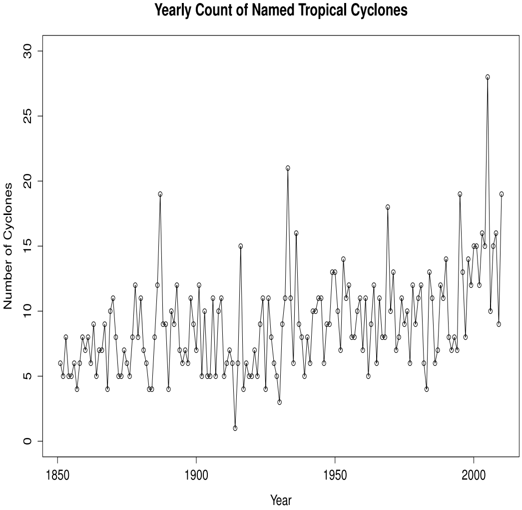
April 29, 2014

Count Time Series

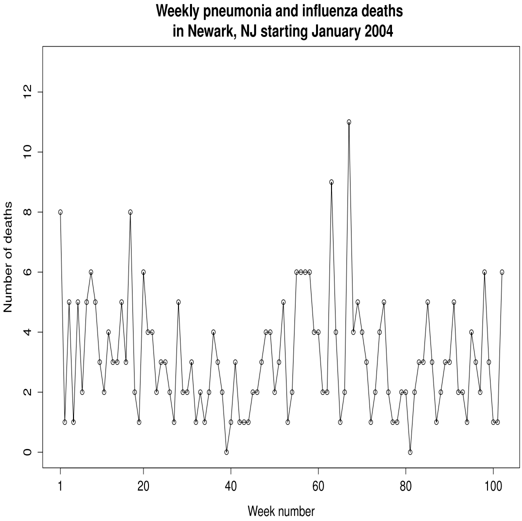
Data taken over time are often integer-valued counts:

- Yearly rare disease cases
- Daily car accidents
- Hourly number of people treated in hospital ERs
- and many others...

Hurricane Data



Flu Death Data



General Issues

Ignoring the count distributional structure (in lieu of Gaussian dynamics) often produces suboptimal inferences, especially when the counts are small.

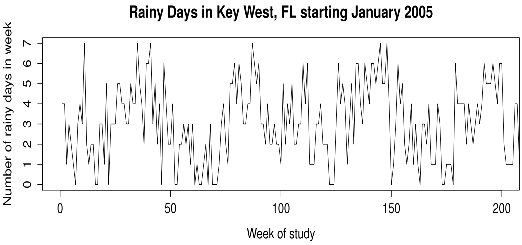
This led to the development of generalized linear models for independent data in the 1970s — tools like Poisson regression emerged.

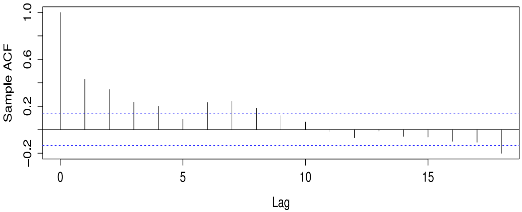
Count time series are often serially correlated.

Cameron and Trivedi (1998) show that one can use standard diagnostic methods (sample correlations, e.g.) to quantify serial correlation in counts.

Non-Gaussian time series versions of count models are needed.

Number of Rainy Days at Key West, Florida





DARMA Methods

DARMA methods (Jacobs and Lewis, late 1970s) construct correlated series having any prescribed marginal distribution π via mixing.

A DAR(1) series $\{X_t\}_{t \geq 0}$ is generated by taking $X_0 \stackrel{D}{=} \pi$ and

$$X_t = V_t X_{t-1} + (1 - V_t) A_t, \quad t \geq 1.$$

Here, $\{A_t\}_{t=1}^{\infty}$ is IID with distribution π and $\{V_t\}_{t=1}^{\infty}$ is an IID sequence of zero/one Bernoulli trials.

Induction shows that $X_t \stackrel{D}{=} \pi$ for all $t \geq 0$.

INARMA Methods

INARMA models mimic ARMA recursions with an appropriate substitute for scalar multiplication to keep the series integer-valued. In INAR(1), for example, obeys

$$X_t = \alpha \circ X_{t-1} + Z_t.$$

Here, $\alpha \in [0, 1]$, $\{Z_t\}$ is an IID non-negative integer-valued count series, and the new multiplication $\circ$ obeys

$$\alpha \circ X = \sum_{i=1}^{X} B_i(\alpha),$$

where $\{B_i(\alpha)\}_{i \geq 1}$ are IID Bernoulli(α).

This can be viewed as a birth/death process with immigration.

DARMA and INARMA Drawbacks

DAR(1) and INAR(1) models cannot have negative autocorrelations as thinning/mixing probabilities must lie in [0,1]. The autocovariance structure is, for $r = \alpha(INAR(1))$ or $r = P(V_t = 1)$ (DAR(1)),

$$\text{Corr}(X_t, X_{t+h}) = r^h.$$

One can define notions of DARMA(p, q) and INARMA(p, q) models. This entails making sense of the causal linear representation

$$X_t = \sum_{\ell=0}^{\infty} \psi_\ell \circ Z_{t-\ell},$$

where $\{Z_t\}$ is IID and count-valued and $\sum_{\ell=0}^{\infty} |\psi_\ell| < \infty$. However, any such model cannot have negative correlations. Also, $\psi_\ell \in [0, 1]$ for all ℓ.

Do Negatively Correlated Counts Series Arise?

Yes:

- Annual Canadian lynx sightings
- North Slope (Alaska) grizzly and polar bear sightings
- Atlantic and Pacific strong hurricanes

We need a new approach.

Renewal Methods

A renewal sequence is a sequence of points, the mth point occurring at

$$L_0 + L_1 + L_2 + \cdots + L_m.$$

Here, $\{L_i\}_{i=1}^{\infty}$ are IID positive integer-valued random variables. L_0 is termed an initial delay. Non-delayed: $L_0 = 0$.

L denotes a random variable whose distribution is that of any L_i for $i \geq 1$. Note: L_0 and L_1 may not have the same distribution.

A renewal sequence of points is made stationary by selecting the distribution of L_0 in a special manner:

$$P[L_0 = k] = E[L]^{-1} P[L > k], \quad k = 0, 1, 2, \ldots .$$

Stationary Renewal Sequences

Given a stationary renewal sequence, set

$$B_t = 1_{[\text{Renewal at time } t]}.$$

Then $\{B_t\}_{t=0}^{\infty}$ is a stationary binary sequence that is autocorrelated in time t:

$$E[B_t] \equiv \frac{1}{E[L]}, \qquad \text{Cov}(B_{t+h}, B_t) = \frac{1}{E[L]}\left(u_h - \frac{1}{E[L]}\right),$$

where u_h is the probability of a renewal at time h in a non-delayed renewal process with lifetimes $\{L_i\}_{i=1}^{\infty}$.

Elementary renewal theorem: $\lim_{h \to \infty} u_h = E[L]^{-1}$.

Superpositioning

Let $\{B_{t,1}\}, \{B_{t,2}\}, \ldots$ be IID copies of $\{B_t\}$. Set

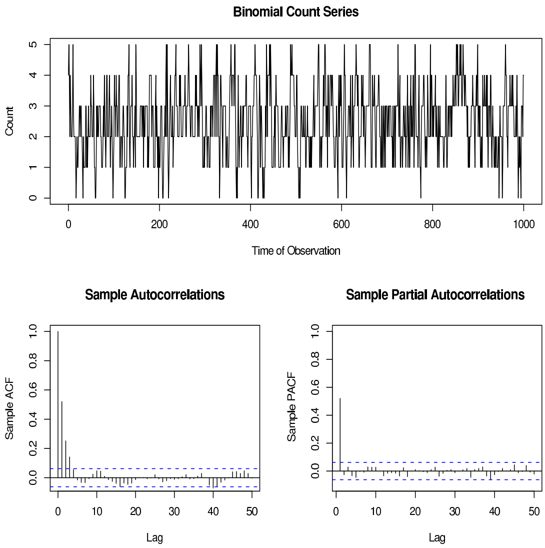
$$X_t = \sum_{\ell=1}^{M} B_{t,\ell}.$$

Then $X_t \sim \text{Bin}(M, E[L]^{-1})$ for each fixed t and

$$\text{Cov}(X_{t+h}, X_t) = \frac{M}{E[L]}\left(u_h - \frac{1}{E[L]}\right), \quad h = 0, 1, 2, \ldots.$$

One will get negative correlation at lag $h > 0$ by taking L such that $u_h < E[L]^{-1}$. This is easy to do.

Realization of a Binomial Count Series with $M = 5$



Poisson Count Series

To get Poisson marginals, set

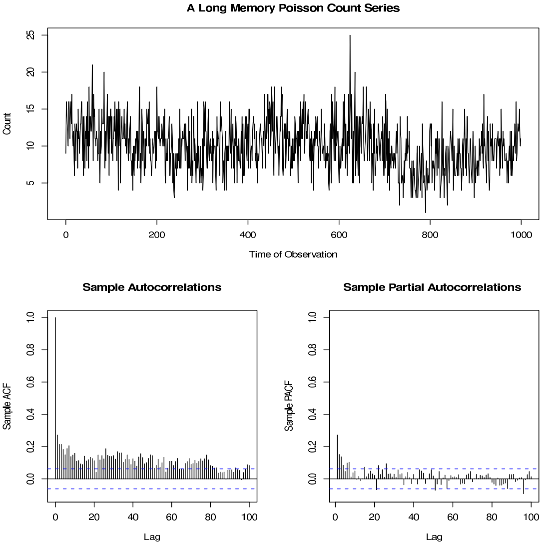
$$X_t = \sum_{\ell=1}^{N_t} B_{t,\ell},$$

where $\{N_t\}$ is IID Poisson with mean $E[N_t] \equiv \lambda$. Then $E[X_t] \equiv \lambda E[L]^{-1}$ and

$$\mathrm{Cov}(X_{t+h}, X_t) = \frac{C(\lambda)}{E[L]} \left(u_h - \frac{1}{E[L]} \right), \qquad h = 1, 2, \ldots.$$

Here, $C(\lambda) = E[\min(N_1, N_2)]$.

Realization of a Long Memory Poisson Count Series



Geometric Count Series

Given $\{B_{t,1}\}, \{B_{t,2}\}, \ldots$ IID stationary renewal count sequences, set

$$X_t = 1 + \sum_{\ell=1}^{N_t} B_{t,\ell},$$

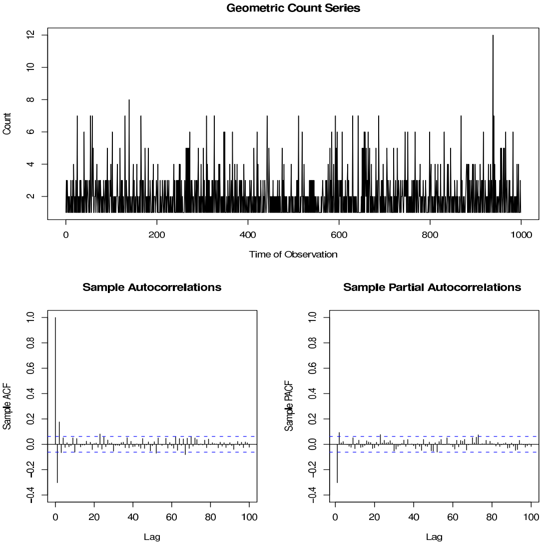
where $\{N_t\}$ are IID zero-modified geometric random variables with success probability α.

Then $\{X_t\}$ has geometric marginals with success probability $p = \alpha E[L]/[\alpha E[L] + (1-\alpha)]$ and

$$\text{Cov}(X_{t+h}, X_t) = \frac{C(\alpha)}{E[L]} \left(u_h - \frac{1}{E[L]} \right),$$

where $C(\alpha) = E[\min(N_1, N_2)]$.

A Realization of a Geometric Count Series



Technical Generalities

Theorem 1: (Cui and Lund, 2009, *Biometrika*) By mixing and superpositioning IID copies of $\{B_t\}$, one can generate stationary series with any prescribed discrete marginal distribution π.

Theorem 2: (Lund and Livsey, 2014, forthcoming). Any pre-specified marginal distribution π supported in $\{0, 1, \ldots\}$ can be achieved via the superposition

$$X_t = c + \sum_{\ell=1}^{N_t} B_{t,\ell}$$

for an appropriately chosen IID count-valued $\{N_t\}$ and integer constant $c \geq 0$. Unfortunately, one cannot have any pre-specified non-negative definite autocovariance.

Technical Generalities (Continued)

Theorem 3: (Cui and Lund, 2009, *Biometrika*). The generated series will have long memory in that

$$\sum_{h=0}^{\infty} |\text{Cov}(X_{t+h}, X_t)| = \infty$$

if and only if $E[L^2] = \infty$.

We used a Pareto lifetime L to generate the long memory Poisson series — specifically, $\alpha = 2.5$ in

$$P(L = k) \propto k^{-\alpha}, \quad k = 1, 2, \ldots.$$

Why is this Cool?

Challenge yourself to devise mechanisms to generate a random pair X and Y, each having a Poisson marginal distribution with the same mean, but with a negative correlation. It's harder than it seems!

For a stationary series $\{X_t\}$ with Poisson marginals, our methods give

$$\text{Corr}(X_{t+1}, X_t) = \frac{C(\lambda)}{\lambda} \left(\frac{u_1 - E[L]^{-1}}{1 - E[L]^{-1}} \right).$$

Now take L as a three point lifetime on $\{1, 2, 3\}$ with the probabilities $P(L = 1) = P(L = 3) = \epsilon$ and $P(L = 2) = 1 - 2\epsilon$ for some small ϵ. Then $u_1 = \epsilon$, $E[L] = 2$, and

$$\text{Corr}(X_{t+1}, X_t) = \frac{C(\lambda)}{\lambda}(2\epsilon - 1).$$

Why is this Cool (Continued)?

Letting $\epsilon \downarrow 0$ gives

$$\mathrm{Corr}(X_{t+1}, X_t) \sim -\frac{C(\lambda)}{\lambda}.$$

This is very close to the most negative correlation possible! Other choices of L are also interesting.

Here,

$$C(\lambda) = \lambda \left(1 - e^{-2\lambda}[I_0(2\lambda) + I_1(2\lambda)]\right);$$

$$I_j(x) = \sum_{n=0}^{\infty} \frac{(x/2)^{2n+j}}{n!(n+j)!}, \quad j = 0, 1.$$

Inference Issues

The renewal model class is very parsimonious, with all parameters evolving from the lifetime L and N_t.

Suppose we want to model the Key West weekly rain day counts. It is natural to pursue models with binomial marginals with $M = 7$.

We will fit the three-parameter geometric mixture lifetime

$$P(L = k) = \xi p_1 (1 - p_1)^{k-1} + (1 - \xi) p_2 (1 - p_2)^{k-1}, \quad k \geq 1.$$

$\xi, p_1, p_2 \in [0, 1], \qquad p_1 \neq p_2$

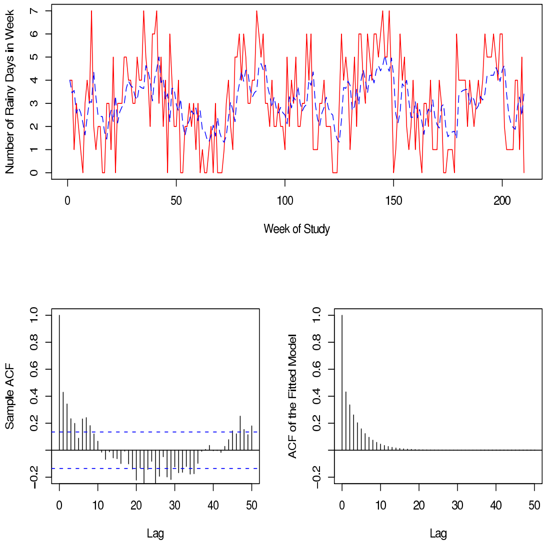
Inference Continued

Our methods estimate ξ, p_1, and p_2 by minimizing the squared sum of one-step-ahead predictions

$$\sum_{t=1}^{n}(X_t - \hat{X}_t)^2,$$

where $\hat{X}_t = P(X_t|1, X_1, \ldots, X_{t-1})$ is the optimal one-step-ahead linear prediction of X_t from past observations. $\hat{X}_t$ depends on ξ, p_1, and p_2.

Key West Data, Jan 2, 2005 – Jan 3, 2009



Current Work

Currently investigated issues include:

1. Periodic renewal models
2. Multivariate renewal models
3. Renewal models with covariates
4. Asymptotic inference