Latent Gaussian Count Time Series Modeling

Stefanos Kechagias (SAS Institute)
jointly with Y. Jia, J. Livsey, R. Lund and V. Pipiras

ISBIS
Piraeus 2018

Outline

1. Motivation
2. Definitions and models
3. Statistical inference
4. Simulation performance and data application

Motivation - Dependent count time series

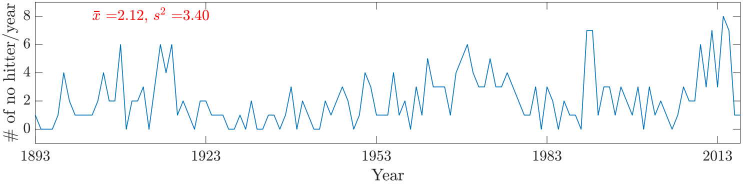
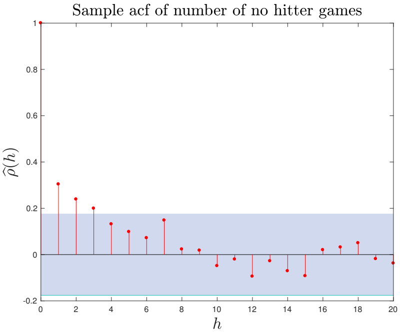
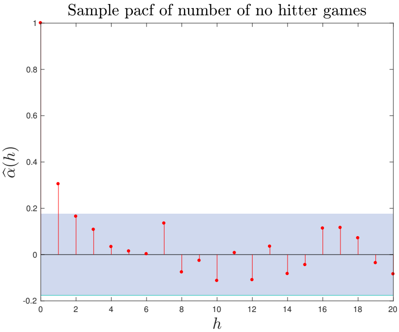


Figure: Yearly counts of no-hitter baseball games from 1875 to 2017.

- Over-dispersed time series of counts
- Negative Binomial, Generalized Poisson or some other distribution?

Motivation - Dependent count time series





- Possible dependence across time

We need count time series (cts) models that allow for flexible dependence structure and can produce any prescribed marginal distribution.

Copula transformation and latent Gaussian variable

- $\{Z_t\}_{t\in\mathbb{Z}}$: stationary, correlated, standard Gaussian series with cdf Φ.
  $\{X_t\}_{t\in\mathbb{Z}}$: stationary cts with desired marginal cdf $F(x) = \mathbb{P}(X_t \leq x)$.

- We model $\{X_t\}$ as

$$X_t = G(Z_t), \quad G(x) = F^{-1}(\Phi(x)), \quad x \in \mathbb{R}, \tag{1}$$

  where F^{-1} is the generalized inverse (quantile function) of F.

- By construction $\{X_t\}$ has marginal cdf F for each t.

How can we associate the dependences structure of $\{Z_t\}$ and $\{X_t\}$?

Linking dependence through acfs

- **Idea:** Try to link the acfs ρ_Z and ρ_X of the two series as

$$\rho_X(h) = \ell(\rho_Z(h))$$

through some function $\ell : [-1, 1] \to [-a, 1]$, $0 < a < 1$.

- ℓ should be $1 - 1$, feasibly computed, and yield large values of a.

- **Solution:** Expand G using *Hermite* polynomials (HP)

$$G(z) = \sum_{k=1}^{\infty} g_k H_k(z),$$

$$H_0 = 1, \quad H_1 = z, \quad H_2 = z^2 - 1, \quad H_3 = z^3 - 3z, \quad H_4 = z^4 - 6z^2 + 3.$$

HP properties

$$H_k(z) = (-1)^k e^{z^2/2} \frac{d^k}{dz^k} e^{-z^2/2},$$

1. HP form an orthogonal basis in $L^2(\mathbb{R}, \phi)$, $G(Z_t) = \sum_{k=1}^{\infty} g_k H_k(Z_t)$
2. $\text{Cov}(H_k(Z_t), H_k(Z_{t+h})) = k! \rho_Z(h)^k$
3. $\text{Cov}(G(Z_t), G(Z_{t+h})) = \sum_{k=0}^{\infty} k! g_k^2 \rho_Z(h)^k$, $\quad g_k = \frac{1}{k!} \mathbb{E}[G(Z_0) H_k(Z_0)]$

Link function & link coefficients

- We associate the acfs ρ_Z and ρ_X of the series $\{Z_t\}$ and $\{X_t\}$ as

$$\rho_X(h) = \sum_{k=1}^{\infty} \frac{k! g_k^2}{\sigma_X^2} \rho_Z(h)^k = \ell(\rho_Z(h)), \quad \ell(u) = \sum_{k=1}^{\infty} \ell_k u^k,$$

where $\ell(\cdot)$ and ℓ_k are called *link* function and *link* coefficients LC.

How flexible is the resulting dependence structure?

Link function properties

- Short memory in Z_t passes on to X_t
- Long memory in Z_t is also inherited in X_t for most marginals.
- $\ell(\cdot)$ yields the largest negative attainable correlation between two variables X_{t_1}, X_{t_2} with the same marginal distribution.

How should we calculate ℓ_k and ℓ?

Calculating LC

Letting $C_n = \mathbb{P}(X_t \leq n)$ and using HP properties we derive

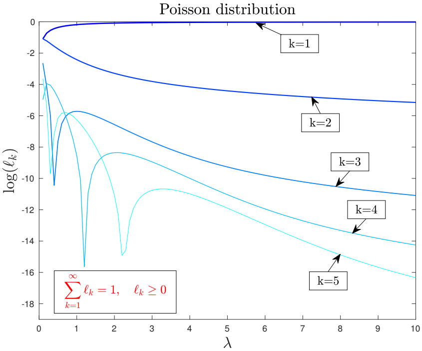
$$g_k = \frac{1}{k!\sqrt{2\pi}} \sum_{n=0}^{\infty} e^{-\Phi^{-1}(C_n)^2/2} H_{k-1}(\Phi^{-1}(C_n)). \quad (2)$$

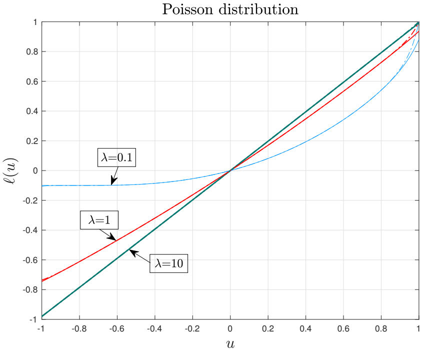
1. (2) converges for processes with finite variance (not obvious).
2. For fairly light-tailed C_n, $C_n \approx 1$ for $n > M$ for small/moderate M.
3. Truncation, HP asymptotics and Stirling's formula yield as $k \to \infty$

$$g_k(k!)^{1/2} \sim \frac{2^{-1/4}}{(\pi k)^{3/4}} \sum_{n=0}^{M-1} e^{-\frac{\Phi^{-1}(C_n)^2}{4}} \cos\left(\Phi^{-1}(C_n)\sqrt{k-1} - \frac{(k-1)\pi}{2}\right),$$

an approximation we found to be accurate even for moderate k.

Plotting LC and LF for Poisson(λ)





1. $\lambda \geq 1$: ℓ_1 carries all the *weight* of the LC (left) and $\ell(u) \approx u$ (right).
2. $\lambda < 1$: *weight* spreads across many LC negative ρ_X are impossible.

Inference

- $\boldsymbol{\theta}$, $\boldsymbol{\eta}$: marginal distribution and latent Gaussian acf parameters

- We approximate the likelihood

$$\mathcal{L}_T(\boldsymbol{\theta}, \boldsymbol{\eta}) = P(X_0 = x_0, X_1 = x_1, \ldots, X_T = x_T)$$

in two ways, through Gaussian likelihood and particle filtering (PF).

- To use PF approximation we write

$$\mathcal{L}_T(\boldsymbol{\theta}, \boldsymbol{\eta}) = \mathbb{P}(X_0 = x_0) \prod_{s=1}^{T} \mathbb{E}_X(w_s(\widehat{Z}_{s|s-1})), \tag{3}$$

where w_s is easily computed from DL quantities and the cdf of X_t.

Connection with HMM

1. When Z_t is an AR(p) process, our model is an HMM.

2. Generate particles Z_t^i and compute weights w_t^i, $i = 1, \dots, N$, using PF sampling algorithms (SIS, SISR, APF).

3. We can then approximate $\mathbb{E}_X[f(\widehat{Z}_{t|t+1})]$ for some function f, and $\mathcal{L}$ as

$$\widehat{\mathbb{E}}_X f(\widehat{Z}_{t+1|t}) = \frac{\frac{1}{N}\sum_{i=1}^N w_t^i f(\widehat{Z}_{t+1}^i)}{\frac{1}{N}\sum_{i=1}^N w_t^i},$$

$$\widehat{\mathcal{L}}_T(x_0, \dots, x_T) = \mathbb{P}(X_0 = x_0) \prod_{s=1}^T \widehat{\mathbb{E}}_X(w_s(\widehat{Z}_{s|s-1})),$$

Simulations-Poisson-AR(1)

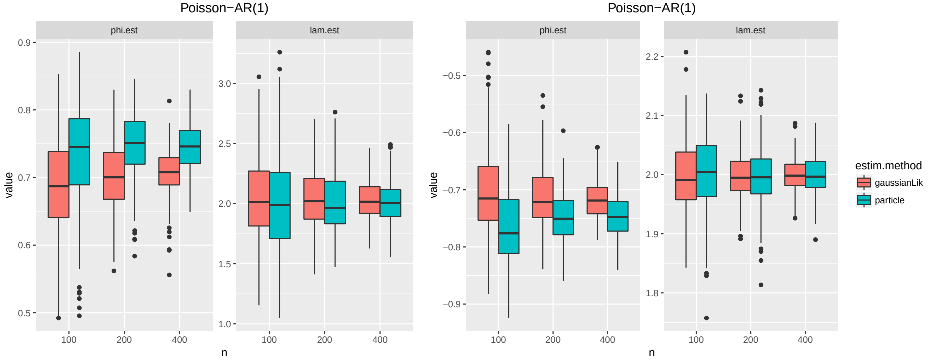
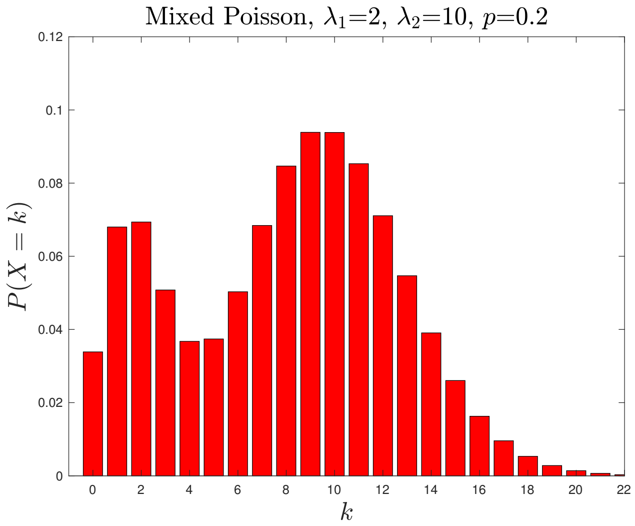


Figure: Estimates from simulated Poisson(2)–AR(1) series with true $\phi = 0.75$ (left) a and $\phi = -0.75$ (right) for sample sizes N= 100, 200, and 400.

Simulations-Mixed Poisson-AR(1)



Simulations-Mixed Poisson-AR(1)

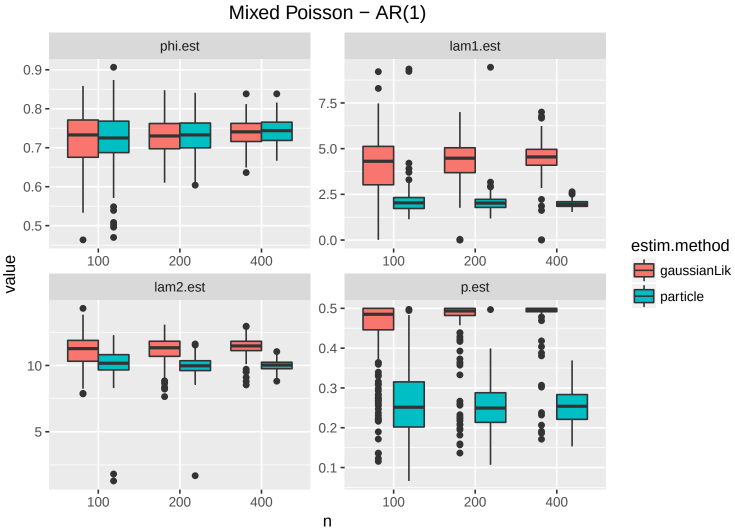


Figure: Estimates from Mixed Poisson(2,10)–AR(1) series with $p = 0.25, \phi = 0.75$

Application to no-hitter data

1. We use Gen. Poisson(η, λ) (overdispersion) and AR(1) (see pacf).
2. We also add two covariates: M_1 the # of games played in a season, and M_2 the height of the pitching mound to the model through

$$\lambda_t = \exp\left(\beta_0 + \beta_1 M_{1,t} + \beta_2 M_{2,t}\right)$$

Parameters	ϕ	β_0	β_1	β_2	η
GL Estimates	0.2665	-1.1496	0.7583	0.0338	0.1679
GL Standard Errors	0.0658	0.9069	0.2173	0.0436	0.0480

Parameters	ϕ	β_1	η
GL Estimates	0.2456	0.4059	0.1212
GL Standard Errors	0.0621	0.0480	0.0416

Summary

Cts model with a latent Gaussian variable

Flexible marginals and dependence structure

Connection with HMM and feasible inference

PIT and residual diagnostics

References

Jia, Y. Kechagias, S. Livsey, J. Lund, R. & Pipiras, V., 'Latent Gaussian Count Time Series', *Preprint*

Pipiras, V. and M. S. Taqqu. *Long-range dependence and self-similarity*. Vol. 45. Cambridge University Press, 2017.

R. Douc, E. Moulines, and D. S. Stoffer. *Nonlinear Time Series: Theory, methods, and applications with R examples.* CRC Press, 2014.

Thank you!