# **GENERALIZED LINEAR MODELS FOR COUNT TIME SERIES**

Nicholas Bosowski; Vinay Ingle

Dept. of ECE, Northeastern University 360 Huntington Avenue, Boston, MA 02115 Email: bosowski.n@husky.neu.edu Email: vingle@ece.neu.edu

# ABSTRACT

In this paper we discuss a class of models for time series of low count data based on the Generalized Linear Model (GLM) approach. Unlike the traditional Auto-Regressive Moving-Average (ARMA) models for continuous Gaussian data, these models capture both the temporal correlation structure and the discrete marginal distribution of count data. We focus on the properties, parameter estimation, and model adequacy aspects for count time series with Poisson or Negative Binomial conditional distributions. The properties and performance of these models are illustrated with synthetic and real data.

# **1. INTRODUCTION**

Time series of counts are obtained in various disciplines whenever a number of events is counted during certain time periods. Examples include the monthly number of car accidents in a region, the weekly number of new cases in epidemiology, the number of transactions at a stock market per minute in finance, or the number of photon arrivals per microsecond in a focal plane array. Modeling low-count time series with ARMA models, which are a popular choice for continuous Gaussian data, leads to inadequate performance. The reason is that ARMA models can capture the temporal correlation structure but fail to refelct the marginal distribution.

Various types of count time series models have been proposed over the years to describe their marginal distribution and correlation structure. In general, these models can be broadly classified into two types: ARMA models based on the notion of thinning and ARMA models based on the GLM approach. Integer ARMA models, which replace multiplication with a "thinning" operation, are discussed in [1]. We focus on count time series models using the GLM approach [2, 3] because they provide a parsimonious manner with which to model count data.

A typical method of estimation is Quasi-Maximum Likelihood Estimation (QMLE) [4], which in the Negative Binomial (NB) case may result in non-stationary models. In this paper, we introduce the Conditional Maximum LikeliDimitris Manolakis

MIT Lincoln Laboratory 244 Wood Street, Lexington, MA 02420–9108 Email: dmanolakis@ll.mit.edu

hood Estimation (CMLE) estimator for the NB, which can be constrained to ensure stationarity. This paper is organized as follows: in Section 2 we introduce count distributions, in Section 3 we discuss how these distributions are used to model correlated time series, in Section 4 we discuss estimation of these models, and in Section 5 we discuss estimation results using synthetic data and give a real world data example.

## 2. DISTRIBUTIONS FOR COUNT DATA

An Independently and Identically Distributed (IID) sequence of Poisson random variables is specified by

$$Y_t \sim \mathsf{Pois}(\lambda) = \frac{\exp(-\lambda)\lambda^y}{y!}, \quad t \ge 1, \ y = 0, 1, 2, \dots$$
(1)

where  $\mathsf{E}(Y_t) = \mathsf{Var}(Y_t) = \lambda$  (equidispersion). Many practical time series exhibit overdispersion, that is,  $\mathsf{Var}(Y_t) > \mathsf{E}(Y_t)$ .

Consider next a Poisson distribution with mean  $\lambda Z$ , where Z > 0 is a random variable with E(Z) = 1 and  $Var(Z) = \sigma_z^2$ . The conditional distribution of  $Y_t$  given Z = zis

$$Y_t | z \sim \mathsf{Pois}(\lambda z) = \frac{\exp(-\lambda z)(\lambda z)^y}{y!}$$
(2)

The unconditional mean and variance are

$$E(Y_t) = E[E(Y_t|z)] = E(\lambda Z) = \lambda E(Z) = \lambda$$
(3)  

$$Var(Y_t) = E[Var(Y_t|Z)] + Var[E(Y_t|Z)]$$
  

$$= \lambda + \lambda^2 \sigma_z^2$$
(4)

We note that the variability of the conditional mean creates overdispersion. We use the normalization E(Z) = 1 to ensure that  $E(Y_t) = \lambda$ . This is the multiplicative analog to the linear (additive) model case  $Y = \lambda + W$  with E(W) = 0.

We usually assume that Z follows a Gamma distribution with shape parameter a > 0 and scale parameter b > 0 defined by

$$f(z;a,b) = \frac{z^{a-1} e^{-z/b}}{\Gamma(a)b^a}, \quad z \ge 0$$
(5)

The mean and variance are E(Z) = ab and  $Var(Z) = ab^2$ . To ensure E(Z) = 1, we set  $a = \nu$  and  $b = 1/\nu$ , which results in  $Var(Z) = 1/\nu$ . The result is the single parameter distribution

$$f(z;\nu) = \frac{\nu^{\nu} z^{\nu-1} e^{-z\nu}}{\Gamma(\nu)}, \quad z \ge 0, \ \nu > 0$$
(6)

This Poisson-Gamma mixture results in the NB distribution, denoted by  $\text{NegBin}(\lambda, \nu)$ , and given by

$$f(y;\nu) = \int_0^\infty f(y|z)f(z;\nu)dz$$
$$= \frac{\Gamma(\nu+y)}{\Gamma(\nu)\Gamma(y+1)} \left(\frac{\nu}{\nu+\lambda}\right)^\nu \left(\frac{\lambda}{\nu+\lambda}\right)^y \quad (7)$$

The mean and variance of which are

$$E(Y) = \lambda, \quad \operatorname{Var}(Y) = \lambda + \frac{1}{\nu}\lambda^2 > \mathsf{E}(Y)$$
 (8)

The NB distribution is widely used to model IID count data with overdispersion.

### 3. LINEAR MODELS FOR COUNT TIME SERIES

In the traditional Gaussian linear models we introduce correlation by filtering a white Gaussian noise process. Since this is impossible for count (integer) time series we follow a different approach. To this end suppose that the conditional distribution of  $Y_t$  given the values of the past observations  $\mathcal{Y}_{t-1} = \{Y_{t-1}, Y_{t-2}, \dots\}$  is given by

$$Y_t | \mathcal{Y}_{t-1} \sim \mathsf{Pois}(\lambda_t) \quad \text{or} \quad Y_t | \mathcal{Y}_{t-1} \sim \mathsf{NegBin}(\lambda_t, \nu)$$
 (9)

where the parameter  $\lambda_t$  changes with time.

In the linear count time series model,  $\lambda_t$  varies according to a hidden or latent process given by

$$\lambda_t = d + a_1 \lambda_{t-1} + b_1 Y_{t-1}, \quad t \ge 1$$
 (10)

Since  $\lambda_t > 0$ , we must have d > 0 and  $\{a_1, b_1\} \ge 0$ . Repeated substitution in (10) shows that

$$\lambda_t = d \frac{1 - a_1^t}{1 - a_1} + a_1^t \lambda_0 + b_1 \sum_{i=0}^{t-1} a_1^i Y_{t-i-1}$$
(11)

The stability of the recursion for  $\lambda_t$  requires that  $|a_1| < 1$ .

To understand the properties of the Poisson linear model, we express (10) as follows

$$Y_{t} = \lambda_{t} + (Y_{t} - \lambda_{t}) = \lambda_{t} + \epsilon_{t}$$
  
=  $d + a_{1}\lambda_{t-1} + b_{1}Y_{t-1} + \epsilon_{t}$   
=  $d + a_{1}(Y_{t-1} - \epsilon_{t-1}) + b_{1}Y_{t-1} + \epsilon_{t}$   
=  $d + (a_{1} + b_{1})Y_{t-1} + (\epsilon_{t} - a_{1}\epsilon_{t-1}),$  (12)

where  $\epsilon_t = Y_t - \lambda_t$  is a white noise process [5] with

$$\mathsf{E}(\epsilon_t) = 0, \ \operatorname{Var}(\epsilon_t) = \mathsf{E}(Y_t), \ \operatorname{Cov}(\epsilon_t, \epsilon_{t+\ell}) = 0, \ \ell > 0$$
(13)

Thus, we obtain the ARMA-like innovations representation

$$(Y_t - m) = (a_1 + b_1)(Y_{t-1} - m) + \epsilon_t - a_1\epsilon_{t-1}$$
(14)

where

$$m = \mathsf{E}(Y_t) = \frac{d}{1 - (a_1 + b_1)} \tag{15}$$

The condition  $0 < a_1 + b_1 < 1$  ensures stationarity and that m > 0.

The variance of the Poisson linear model is

$$\operatorname{Var}(Y_t) = \frac{1 - (a_1 + b_1)^2 + b_1^2}{1 - (a_1 + b_1)^2} m \tag{16}$$

Since  $Var(Y_t) \ge E(Y_t)$ , with equality when  $b_1 = 0$ , the marginal distribution of  $Y_t$  is not Poisson.

The marginal mean of the NB linear model is identical with that of the Poisson model. However, the marginal variance of the NB model is larger than that of the Poisson model

$$\operatorname{Var}(Y_t) = \frac{1 - (a_1 + b_1)^2 + b_1^2}{1 - (a_1 + b_1)^2 - \frac{b_1^2}{\nu}} \left(m + \frac{m^2}{\nu}\right)$$
(17)

The Auto-Covariance Function (ACVF) function  $C_Y(\ell) = Cov(Y_t, Y_{t+|\ell|})$  of the NB linear model for  $|\ell| \ge 1$  is given by

$$C_Y(\ell) = b_1 \frac{1 - a_1(a_1 + b_1)}{1 - (a_1 + b_1)^2 - \frac{b_1^2}{\nu}} \left(m + \frac{m^2}{\nu}\right) (a_1 + b_1)^{|\ell| - 1}$$
(18)

The serial correlation of the Poisson and NB models is the same. We note that as  $\nu \to \infty$ ,  $Var(Z_t) = (1/\nu) \to 0$  and the NB distribution becomes the Poisson distribution. To model negatively correlated data, the log-linear model found in [6,7] can be considered.

#### 4. MODEL PARAMETER ESTIMATION

Consider the general linear model

$$\lambda_t = d + \sum_{i=1}^{q} a_i \lambda_{t-i} + \sum_{j=1}^{p} b_j Y_{t-j}$$
(19)

The log-likelihood function of the linear NB model is

$$l(\boldsymbol{\theta}, \nu) = \sum_{t=1}^{N} \left[ \log \frac{\Gamma(y_t + \nu)}{\Gamma(y_t + 1)\Gamma(\nu)} + \nu \log \frac{\nu}{\nu + \lambda_t(\boldsymbol{\theta})} + y_t \log \frac{\lambda_t(\boldsymbol{\theta})}{\nu + \lambda_t(\boldsymbol{\theta})} \right]$$
(20)

**Table 1**. Results of model estimation for an aggregate set of 1000 time series of length 1000 for each model. The first row for each model shows the mean of the estimator, and the second shows the standard errors. The last row shows how many of the unconstrained estimates of  $\hat{\nu}_2$  were outside the region of stationarity.

Model	QMLE						CMLE				Misc.	
	$\hat{d}$	$\hat{b}_1/\hat{b}_3$	$\hat{a}_1/\hat{b}_2$	$\hat{\nu}_1$	$\hat{\nu}_2^{uc}$	$\hat{\nu}_2^c$	$\hat{b}_0$	$\hat{b}_1/\hat{b}_3$	$\hat{a}_1/\hat{b}_2$	$\hat{\nu}$	$ u^* $	$\hat{\nu}_{uc} \le \nu^*$
$(b_0, b_1, a_1, \nu) =$	0.54	0.49	0.40	1.54	1.52	1.70	0.55	0.49	0.39	1.52	1.32	198
(0.50, 0.50, 0.40, 1.52)	0.11	0.06	0.06	0.15	0.14	0.89	0.17	0.05	0.06	0.11		
$(b_0, b_1, a_1, \nu) =$	0.53	0.50	0.39	6.48	6.35	6.45	0.52	0.50	0.39	6.41	1.32	0
(0.50, 0.50, 0.40, 6.32)	0.09	0.04	0.04	1.03	0.78	0.79	0.09	0.03	0.04	0.73		
$(b_0, b_1, a_1, \nu) =$	1.09	0.69	0.19	2.80	2.77	3.15	1.07	0.70	0.19	2.80	2.58	278
(1.00, 0.70, 0.20, 2.78)	0.18	0.05	0.05	0.28	0.25	2.35	0.17	0.04	0.05	0.23		
$(b_0, b_1, a_1, \nu) =$	1.07	0.70	0.19	7.73	7.60	7.69	1.03	0.70	0.20	7.65	2.58	1
(1.00, 0.70, 0.20, 7.58)	0.17	0.04	0.05	0.94	0.79	0.80	0.14	0.04	0.04	0.73		
$(b_0, b_1, b_2,  u) =$	1.01	0.49	0.36	3.13	3.07	3.28	1.00	0.49	0.36	3.11	2.90	141
(1.00, 0.50, 0.40, 3.10)	0.17	0.05	0.06	0.32	0.28	0.84	0.17	0.04	0.06	0.22		
$(b_0, b_1, b_2, \nu) =$	1.01	0.49	0.37	8.04	7.88	8.01	1.00	0.49	0.37	7.99	2.90	1
(1.00, 0.50, 0.40, 7.90)	0.16	0.04	0.05	1.11	0.85	0.86	0.16	0.03	0.05	0.74		

where  $\theta = (d, a_1, \dots, a_q, b_1, \dots, b_p)$  and  $\lambda_t(\theta)$  Specifies the mean at time t. Maximization is performed in the (constrained) parameter spaces

$$\Theta_{\rm L} = \{ d > 0, a_i \ge 0, b_j \ge 0, \sum_i a_i + \sum_j b_j < 1 \}$$

In practice, under the mixed Poisson model assumption, it is preferable to estimate  $\theta$  by minimizing the quasi-likelihood

$$l(\boldsymbol{\theta}) = \sum_{t=1}^{N} \left\{ y_t \log \lambda_t(\boldsymbol{\theta}) - \lambda_t(\boldsymbol{\theta}) \right\}$$
(21)

under the constraints imposed upon the parameters of each model. Maximization of (21) under a conditional Poisson distribution is CMLE of the Poisson linear model, which is discussed in [3]. Under the conditional NB distribution maximization of (21) is QMLE of the NB linear model, discussed in [4, 8]. The dispersion parameter  $\nu$  is then estimated independently. In [4] the authors suggest the use of either

$$\hat{\nu} = \left(\frac{1}{n} \sum_{i=1}^{N} \frac{\left[(Y_t - \hat{\lambda}_t)^2 - \hat{\lambda}_t\right]}{\hat{\lambda}_t^2}\right)^{-1}$$
(22)

which is a moment based estimator, or

$$N - M = \sum_{t=1}^{N} \frac{(Y_t - \hat{\lambda}_t)^2}{\hat{\lambda}_t (1 + \frac{\hat{\lambda}_t}{\nu})}$$
(23)

which is based on statistics of the NB distribution and solved for  $\nu$ , and M represents the dimension of  $\theta$ . While this methodology is convenient, it suffers in that it can result in estimates corresponding to non-stationary models. To see this we re-write (19) in an ARMA-like representation

$$Y_t = d - \sum_{i=1}^{s} (a_i + b_i) Y_{t-k} + \epsilon_t + \sum_{j=1}^{q} a_j \epsilon_{t-k} \quad (24)$$

where  $s = \max(p, q)$ ,  $a_i = 0$ , i > q,  $b_i = 0$ , i > p. Equation (24) can be equivalently re-written as an infinite moving-average process from which we find

$$\operatorname{Var}(Y_t) = \left(1 + \sum_{k=1}^{\infty} \psi_k^2\right) \operatorname{Var}(\epsilon_t)$$
(25)

Under the assumption of an NB conditional distribution we observe

$$\operatorname{Var}(\epsilon_{t}) = \operatorname{Var}\left[\mathsf{E}\left(\epsilon_{t}|\mathcal{Y}_{t-1}\right)\right] + \mathsf{E}\left[\operatorname{Var}\left(\epsilon_{t}|\mathcal{Y}_{t-1}\right)\right]$$
$$= m + \frac{\mathsf{E}\left(\lambda_{t}^{2}\right)}{\nu}$$
(26)

Using (26) it can be shown that

$$\operatorname{Var}(Y_t) = \frac{\left(\nu m + m^2\right) \left(1 + \sum_{k=1}^{\infty} \psi_k^2\right)}{\nu - \sum_{k=1}^{\infty} \psi_k^2}$$
(27)

from which we observe that

$$\nu > \sum_{k=1}^{\infty} \psi_k^2 = \nu^*$$
(28)

must be true to ensure  $Var(Y_t)$  exists. QMLE suffers in that by separating estimation of  $\theta$  and  $\nu$  it loses the ability to simultaneously constrain the parameter space. Instead we suggest direct maximization of (20), which is CMLE of the NB linear model.



**Fig. 1**. Results of one step estimation on the Nuclear Tests data set. The top plots shows the data; the bottom left plot zooms in on the last 50 points and shows a single additional point marked with a red x; the bottom right plot shows Normal(red), Poisson(Blue), and NB(green) one-step predictive distributions.

## 5. ESTIMATION RESULTS

Table 1 shows estimation results using QMLE and CMLE for several linear NB models. For each model two different values of  $\nu$  were chosen { $\nu^* + .5$ ,  $\nu^* + 5$ } to investigate performance near and far from the stationarity boundary. Results were generated by simulating 1000 time series of 1000 observations and using both the QMLE and CMLE algorithms to perform parameter estimation. Estimation of  $\nu$ for QMLE was performed using both (22), denoted by  $\nu_1$ , and (23), denoted by  $\nu_2$ . Additionally, evaluation of  $\hat{\nu}_2$  was performed both constrained and unconstrained. The estimator  $\hat{\nu}$ for CMLE was constrained as well.

The first row of each model in Table 1 shows the mean estimators whereas the second row reports the standard errors. Comparing QMLE and CMLE for  $\hat{\theta}$  shows that both estimators perform equally well in terms of both bias and convergence. This is not the case however, for estimating  $\nu$ . When  $\nu$  is far away from the boundary constraining the solution of  $\hat{\nu}_2$  has little to no effect on the bias and convergence, and the estimator has similar bias and standard error to the CMLE estimator.

When  $\nu$  is close to the stationarity boundary however, constraining  $\hat{\nu}_2$  drastically biases the estimator and increases the standard error. We do not observe this phenomenon with the CMLE estimator of  $\nu$  since the constrained space is always considered. For each model where  $\nu$  is near the boundary a sizable portion of the unconstrained estimators give non-stationary results, reflected in the last column of Table 1. The moment based estimator  $\nu_1$  performs worse than the constrained CMLE estimator and additionally cannot be constrained. Additionally, as observed in [4] this estimator performs worse than the unconstrained version  $\hat{\nu}_2$ , therefore we do not suggest its use.

Another advantage CMLE has over QMLE is that it naturally provides standard errors on  $\hat{\nu}$  via the conditional likeli-

hood information matrix. If one uses QMLE, a bootstrapping technique using the estimated model must be performed to find such bounds. This can be problematic when the estimates are either non-stationary if the estimator is unconstrained or biased if they are.

Figure 1 shows an application of the linear models to a test set counting the number of nuclear tests performed monthly by the United States between 1961 and 1992 [9]. We observe that due to the low count values of the data, the one-step predictive distribution using an ARMA model, shown in red, fits poorly. The Poisson model in blue, and the NB in green both appear to fit the data much better.

#### 6. CONCLUSION

This paper has shown that while QMLE and CMLE can both be used for estimation of NB linear models, CMLE has the benefit of being able to naturally constrain the parameter space. Additionally, the CMLE estimator produces standard errors on the estimate of  $\nu$ , whereas the QMLE must resort to bootstrapping. This can suffer from either non-stationary model estimates if the estimator of  $\nu$  is left unconstrained or bias if the estimator is constrained. We have additionally shown how we can find the marginal variance of any NB linear model, which is important in understanding the model dynamics for higher order and non-trivial models.

# Acknowledgment

This work is sponsored by the Air Force Research Laboratory under Air Force Contract #FA8702-15-D-0001. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the United States Government.

# 7. REFERENCES

- M. G. Scotto, C. H. Weiß, and S. Gouveia, "Thinningbased models in the analysis of integer-valued time series: a review," *Statistical Modelling*, p. 1471082X15584701, 2015.
- [2] B. Kedem and K. Fokianos, *Regression models for time series analysis*. John Wiley & amp; Sons, 2005, vol. 488.
- [3] K. Fokianos, A. Rahbek, and D. Tjøstheim, "Poisson autoregression," *Journal of the American Statistical Association*, vol. 104, no. 488, pp. 1430–1439, 2009.
- [4] V. Christou and K. Fokianos, "Quasi-likelihood inference for negative binomial time series models," *Journal* of *Time Series Analysis*, vol. 35, no. 1, pp. 55–78, 2014.
- [5] K. Fokianos, "Count time series models," *Time Series–Methods and Applications*, no. 30, pp. 315–347, 2012.
- [6] K. Fokianos and D. Tjøstheim, "Log-linear Poisson autoregression," *Journal of Multivariate Analysis*, vol. 102, no. 3, pp. 563–578, 2011.
- [7] R. Douc, P. Doukhan, and E. Moulines, "Ergodicity of observation-driven time series models and consistency of the maximum likelihood estimator," *Stochastic Processes and their Applications*, vol. 123, no. 7, pp. 2620–2647, 2013.
- [8] V. Christou and K. Fokianos, "On count time series prediction," *Journal of Statistical Computation and Simulation*, vol. 85, no. 2, pp. 357–373, 2015.
- [9] [Online]. Available: http://www.ga.gov.au/oracle/nuclear-explosion-query.jsp