Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v10n3p654

Modeling football data using a GQL algorithm based on higher ordered covariances
By Jowaheer, Mamode Khan, Sunecher

Published: 15 November 2017

Electronic Journal of Applied Statistical Analysis
Vol. 10, Issue 03, November 2017, 654-665
DOI: 10.1285/i20705948v10n3p654

Modeling football data using a GQL algorithm based on higher ordered covariances

Jowaheer V.[a], Mamode Khan N.*[b], and Sunecher Y.[c]

[a] *University of Mauritius, Department of Mathematics, Reduit*
[b] *University of Mauritius, Department of Economics and Statistics, Reduit*
[c] *University of Technology Mauritius, Department of Accounting and Finance, Pointe-Aux-Sables*

Published: 15 November 2017

This paper deals with the modeling of the first and second half number of football goals using a bivariate integer-valued first-order autoregressive model (BINAR(1)) with Negative Binomial (NB) innovations defined under time-dependent moments. The main novelty of the paper is the estimation of the regression and over-dispersion parameters via a generalized quasi-likelihood (GQL) function wherein some components of the auto-covariance structure are computed through the 'working' multivariate normality assumption. The model is assessed on the Arsenal Football Club data from the period 2010 to 2016. The result of the study has revealed some important findings that illustrate the downfall of the club in the recent years as regards to the Premier League matches.

keywords: Bivariate, Negative Binomial, Over-Dispersion, Non-Stationarity.

1 Introduction

The modeling of football goals is an interesting topic of research for both sports people and statisticians. In literature, the research papers in this field mainly focus on the modeling of goals scored by opponent teams (see Maher, 1982, Baxter and Stevenson, 1998, Karlis and Ntzoufras, 2000, 2003, Baio and Blangiardo, 2010 and the references

*Corresponding author: n.mamodekhan@uom.ac.mu

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

therein). Initially, the models were built on the assumption that the goals by the teams are independent until Karlis and Ntzoufras (2000, 2003) proposed more sophisticated models in the form of the bivariate Poisson and inflated bivariate Poisson that assume that the goals are inter-related. Current studies in this field (Groll and Abedieh, 2013, Groll et al., 2015, Karlis and Ntzoufras, 2010 and Louzada et al., 2014) focus mainly on the prediction of the football outcomes but yet there is no paper that analyzes the series of first and second half number of goals scored by a football team.

Admittedly, in any football league, the number of goals scored or the goal difference is an important performance indicator that determines the strength of the football team. Such figures may also determine the winner of the league in case of an ex-aequo, to determine the ranking of the team in the league table and ultimately its possibility of participating in other prominent competitions such as the UEFA Champions League. It is quite rationale to believe that the first half and second half performance of a football team is highly inter-related and the second half performance depends extensively on the first half scores.

In this paper, we propose a bivariate time series model based on an integer-valued autoregressive (BINAR(1)) process to analyze the series of first half and second half number of goals scored by a football team. In the proposed BINAR(1) model, the cross-relation between the series is assumed to be induced by some jointly distributed innovation terms. As at date, several BINAR(1) processes have been developed to model correlated series of counts under various probability functions and under different cross-correlation structures. Originally, Pedeli and Karlis (2009, 2011) proposed a simple and constrained BINAR(1) model with Poisson innovations where the inter-relation between the series was due to the correlated innovations only but this model was developed only under stationary moments. Recently, Mamode Khan et al. (2016) proposed an extension of this model by considering that the two series are influenced by some common time-varying explanatory variables and this ultimately results into non-stationary BINAR(1) with Poisson marginals. In the same trend, Sunecher et al. (2017) developed a non-stationary BINAR(1) process with Negative Binomial (NB) innovations wherein both series are over-dispersed under different levels of over-dispersion. Since usually the number of goals scored express huge variability across time, this paper makes reference to the latest BINAR(1) model by Sunecher et al. (2017) to analyze the first and second half performance. However, the above BINAR(1) process is re-formulated as per the assumptions illustrated in the references pertaining to football studies. Moreover, a similar generalized quasi-likelihood (GQL) estimation approach described in Sunecher et al. (2017) is utilized to estimate the mean or regression effects and the dispersion coefficients, since the estimation method requires only the marginal properties of the bivariate series (see Mamode Khan et al., 2016; Sunecher et al., 2017). It is worth mentioning that in the estimation of the over-dispersion coefficients for both series, the auto-covariance expressions depend on some high-ordered moments. From Prentice and Zhao (1991) and Sunecher et al. (2017), these are computed using the 'working' multivariate normality assumption but we demonstrate in Section 3 that in the presence of some realistic football assumptions, some of the off-diagonal high-ordered entries have simpler expressions. As for the covariate selection, in this study, the Home and Away status, the number

of new players recruited in the transfer window, the number of players sold, retired or loaned to other clubs, the number of inter-matches played between two Premier League matches and the number of players injured in the inter-matches other than the normal league are taken into consideration.
The organization of the paper is as follows: In Section 2, the BINAR(1) model for football data is introduced and the cross-covariances are derived. In Section 3, we estimate the unknown parameters by using two sets of GQL equations. This Section is followed by the derivation of the forecasting equations and a case study on Arsenal football data is presented in Section 5. The conclusion is provided in Section 6.

2 The Bivariate INAR(1) Time Series Process

McKenzie (1986) developed an integer-valued autoregressive process of order 1 (INAR(1)) that relates count observation at a specified time point with its previous lagged observation and with an innovation or residual term. Pedeli and Karlis (2009, 2011) proposed a direct extension of this classical INAR(1) by considering two INAR(1) series that are interrelated through the jointly distributed innovation terms. However, this extension was restricted only to series that exhibit stationary moments.
In view of the number of applications that involve time-variant moments or explanatory variables, Mamode Khan et al. (2016) derived a novel BINAR(1) model wherein the series were non-stationary Poisson marginals with bivariate Poisson distributed innovation terms. However, this model could not accommodate over-dispersion. In a recent research, Sunecher et al. (2017) proposed a non-stationary BINAR(1) process with NB innovations defined under time-dependent moments at different levels of over-dispersion. This section reviews this BINAR(1) model but under some new assumptions that relate only to football studies. It is quite reasonable that the number of goals in the first half of the t^{th} league match is not related with the second half of the $(t+h)^{th}$ league match or vice versa. Hence, by denoting $Y_t^{[k]}$ as the number of goals in the k^{th} half of the t^{th} league match, for $k = 1, 2$, then

$$\text{Cov}(Y_t^{[1]}, Y_{t+h}^{[2]}) = 0, \quad h > 0. \tag{1}$$

From an extension of McKenzie (1986), let:

$$Y_t^{[1]} = \alpha_1 * Y_{t-1}^{[1]} + R_t^{[1]}. \tag{2}$$

$$Y_t^{[2]} = \alpha_2 * Y_{t-1}^{[2]} + R_t^{[2]}. \tag{3}$$

where α_k are randomly Beta-distributed parameters in the interval [0,1] with $\alpha_k \sim \text{Beta}(\frac{\rho_k}{c_k}, \frac{1-\rho_k}{c_k})$ (see McKenzie, 1986) and $'*'$ is the binomial thinning operator (Aly and Bouzar, 2005; Bourguignon and Vasconcellos, 2015; McKenzie, 1988; Silva and Oliveira, 2004; Steutel and Van Harn, 1979; Weiβ, 008b) such that $\alpha_k * Y_{t-1}^{[k]} \mid Y_{t-1}^{[k]} \sim \text{Binomial}(Y_{t-1}^{[k]}, \alpha_k)$, while $R_t^{[k]} \sim \text{NB}(\frac{1}{c_k^*}, c_k^*(\mu_t^{[k]} - \rho_k \mu_{t-1}^{[k]}))$ and $Y_t^{[k]} \sim \text{NB}(\frac{1}{c_k}, c_k \mu_t^{[k]})$ (see

Sunecher et al., 2017), where for any random variable Y, denoting $Y \sim NB(\frac{1}{c}, c\mu)$, it signifies that $f_Y(y)$ has a NB distribution of the form

$$f_Y(y) = \frac{\Gamma(c^{-1}+y)}{\Gamma(c^{-1})y!}\left(\frac{1}{1+c\mu}\right)^{c^{-1}}\left(\frac{c\mu}{1+c\mu}\right)^{y}, \quad \nu \geq 0, \quad c > 0, \tag{4}$$

with $E(Y) = \mu$, $Var(Y) = \mu + c\mu^2$ and c indicates the over-dispersion parameter.
The sequence $\{R_t^{[k]}\}_{t=1}^{T}$ is mutually independent for a specific k and the pair of $(Y_{t-1}^{[k]}, R_t^{[k]})$ are independent across all $t = 1, \ldots, T$ and $k = 1, 2$. We allow the pair $(R_t^{[1]}, R_t^{[2]})$ be inter-related such that

$$\text{Corr}(R_t^{[1]}, R_{t'}^{[2]}) = \begin{cases} \rho_{12,t} & t = t', \\ 0 & t \neq t'. \end{cases} \tag{5}$$

Based on the binomial thinning properties and using the moments of the Beta distribution such that $E(\alpha_k) = \rho_k$ and $\text{Var}(\alpha_k) = \frac{\rho_k(1-\rho_k)c_k}{1+c_k}$, we obtain

$$\begin{aligned} E(Y_t^{[k]}) &= E_{Y_{t-1}^{[k]}} E_{\alpha^{[k]}} E(\alpha^{[k]} * Y_{t-1}^{[k]} | Y_{t-1}^{[k]}, \alpha^{[k]}) + E(R_t^{[k]}) \\ &= E(\rho_k Y_{t-1}^{[k]}) + E(R_t^{[k]}) \\ &= \mu_t^{[k]}. \end{aligned} \tag{6}$$

$$\begin{aligned} \text{Var}(Y_t^{[k]}) &= \text{Var}(\alpha^{[k]} * Y_{t-1}^{[k]}) + \text{Var}(R_t^{[k]}) \\ &= \text{Var}_{\alpha^{[k]}}[E(\alpha^{[k]} * Y_{t-1}^{[k]} | Y_{t-1}^{[k]}, \alpha^{[k]})] + E_{\alpha^{[k]}}[\text{Var}(\alpha^{[k]} * Y_{t-1}^{[k]} | Y_{t-1}^{[k]}, \alpha^{[k]})] + \text{Var}(R_t^{[k]}) \\ &= \text{Var}_{\alpha^{[k]}}(\alpha^{[k]}\mu_{t-1}^{[k]}) + E_{\alpha^{[k]}}[\alpha^{[k]}(1-\alpha^{[k]})\mu_{t-1}^{[1]} + {\alpha^{[k]}}^2(\mu_{t-1}^{[k]} + c_k {\mu_{t-1}^{[k]}}^2)] + \text{Var}(R_t^{[k]}) \\ &= {\mu_{t-1}^{[k]}}^2[\frac{\rho_k(1-\rho_k)c_k}{1+c_k}] + \rho_k\mu_{t-1}^{[k]} + c_k{\mu_{t-1}^{[k]}}^2[\frac{\rho_k(1-\rho_k)c_k}{1+c_k} + \rho_k^2] \\ &+ (\mu_t^{[k]} - \rho_k\mu_{t-1}^{[k]}) + c_k^*(\mu_t^{[k]} - \rho_k\mu_{t-1}^{[k]})^2. \end{aligned} \tag{7}$$

By comparing $\text{Var}(Y_t^{[k]}) = \mu_t^{[k]}(1 + c_k\mu_t^{[k]})$ with equation (7), it is deduced that $c_k^* = \frac{c_k({\mu_t^{[k]}}^2 - \rho_k {\mu_{t-1}^{[k]}}^2)}{(\mu_t^{[k]} - \rho_k\mu_{t-1}^{[k]})^2} > 0$. In the above, it is assumed $\mu_t^{[k]} = \exp(\boldsymbol{x}_t'\boldsymbol{\beta}^{[k]})$ with $\boldsymbol{x}_t = [x_{t1}, x_{t2}, \ldots, x_{tp}]'$ and $\boldsymbol{\beta}^{[k]} = [\beta_1^{[k]}, \beta_2^{[k]}, \ldots, \beta_p^{[k]}]$.
Note also,

$$\text{Corr}(R_t^{[1]}, R_t^{[2]}) = \text{Corr}(Y_t^{[1]}, R_t^{[2]}) = \text{Corr}(Y_t^{[2]}, R_t^{[1]}) \tag{8}$$

Hence, using the above it can easily be shown that

$$\text{Corr}(Y_t^{[k]}, Y_{t+h}^{[k]}) = \rho_k^h \frac{\sqrt{(\mu_t^{[k]} + c_k(\mu_t^{[k]})^2)}}{\sqrt{\mu_{t+h}^{[k]} + c_k(\mu_{t+h}^{[k]})^2}} \tag{9}$$

and

$$\text{Corr}(Y_t^{[1]}, Y_{t+h}^{[2]}) = \begin{cases} \frac{\rho_1\rho_2\text{Cov}(Y_{t-1}^{[1]}, Y_{t+h-1}^{[2]}) + \rho_{12,t}\sqrt{\lambda_t^{[1]} + c_1^*{\lambda_t^{[1]}}^2}\sqrt{\lambda_{t+h}^{[2]} + c_2^*{\lambda_{t+h}^{[2]}}^2}}{\sqrt{\mu_t^{[1]} + c_1{\mu_t^{[1]}}^2}\sqrt{\mu_{t+h}^{[2]} + c_2{\mu_{t+h}^{[2]}}^2}} & h = 0, \\ 0 & h \neq 0 \end{cases} \tag{10}$$

3 Generalized Quasi-Likelihood (GQL)

This section introduces the two GQL equations to estimate the regression and over-dispersion parameters respectively for the two series. From Mamode Khan et al. (2016) and Sunecher et al. (2017), the QL equations generally consists of the derivative structure, auto-covariance components and most importantly the score function. As for the regression parameter $\boldsymbol{\beta}^{[k]}$, the estimating function is expressed as

$$\boldsymbol{D}_{\boldsymbol{\beta}}^{'}{\boldsymbol{\Sigma}_{\boldsymbol{\beta}}}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}) = 0, \tag{11}$$

where $\boldsymbol{Y} = [\boldsymbol{Y}^{[1]}, \boldsymbol{Y}^{[2]}]'$ with $\boldsymbol{Y}^{[k]} = [Y_1^{[k]}, Y_2^{[k]}, \ldots, Y_T^{[k]}]'$ and $\boldsymbol{\mu}$ is the corresponding mean vector.
The derivative part is a $(2T \times 2p)$ block diagonal $[\boldsymbol{D}^{[1]}, \boldsymbol{D}^{[2]}]$ with the entries in $\boldsymbol{D}^{[k]}$ specified as $\frac{\partial \mu_t^{[k]}}{\partial \beta_j^{[k]}} = exp(x_{tj}^{'}\beta_j^{[k]})x_{tj}^{'}$.
As for the auto-covariance components,

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \begin{pmatrix} \text{Var}(Y^{[1]}) & \text{Cov}(Y^{[1]}, Y^{[2]}) \\ \text{Cov}(Y^{[2]}, Y^{[1]}) & \text{Var}(Y^{[2]}) \end{pmatrix} \tag{12}$$

with the entries in $\text{Var}(Y^{[k]})$ specified as

$$\text{Cov}(Y_t^{[k]}, Y_{t+h}^{[k]}) = \begin{cases} \mu_t^{[k]} + c_k(\mu_t^{[k]})^2 & h = 0, \\ \rho_k^h(\mu_t^{[k]} + c_k(\mu_t^{[k]})^2) & h \neq 0 \end{cases} \tag{13}$$

where

$$\hat{\rho}_k = \frac{T\sum_{t=1}^{T-1}\tilde{Y}_t^{[k]}\tilde{Y}_{t+1}^{[k]}}{[\sum_{t=1}^{T}(\tilde{Y}_t^{[k]})^2][\sum_{t=2}^{T}\sqrt{\frac{(\mu_t^{[k]} + c_k(\mu_t^{[k]})^2)}{\mu_{t+1}^{[k]} + c_k(\mu_{t+1}^{[k]})^2}}]} \tag{14}$$

with $\tilde{Y_t}^{[k]} = \frac{Y_t^{[k]} - \mu_t^{[k]}}{\sqrt{\mu_t^{[k]} + c_k{\mu_t^{[k]}}^2}}$ and $\rho_{12,t}$ is obtained using Equation (10) as follows:

$$\hat{\rho}_{12,t} = \frac{\tilde{\text{Cov}}(Y_t^{[1]}, Y_t^{[2]}) - \hat{\rho}_1\hat{\rho}_2\tilde{\text{Cov}}(Y_{t-1}^{[1]}, Y_{t-1}^{[2]})}{\sqrt{\hat{\lambda}_t^{[1]} + \hat{c}_1^*(\hat{\lambda}_t^{[1]})^2}\sqrt{\hat{\lambda}_t^{[2]} + \hat{c}_2^*(\hat{\lambda}_t^{[2]})^2}} \tag{15}$$

$\tilde{\text{Cov}}(Y_t^{[1]}, Y_t^{[2]})=\frac{1}{T}\sum_{t=1}^{T}(y_t^{[1]}-\hat{\mu}_t^{[1]})(y_t^{[2]}-\hat{\mu}_t^{[2]})$.
Re-arranging and solving Equation (11) yields a Newton Raphson iterative scheme of the form

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}_{r+1}^{[1]} \\ \hat{\boldsymbol{\beta}}_{r+1}^{[2]} \end{pmatrix} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_{r}^{[1]} \\ \hat{\boldsymbol{\beta}}_{r}^{[2]} \end{pmatrix} + [\boldsymbol{D}_{\boldsymbol{\beta}}^{'}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}{}^{-1}\boldsymbol{D}_{\boldsymbol{\beta}}]_r^{-1}[\boldsymbol{D}_{\boldsymbol{\beta}}{}^{'}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}{}^{-1}(\boldsymbol{Y}-\boldsymbol{\mu})]_r \tag{16}$$

Note that under mild regularity and asymptotically conditions, these estimates are consistent and it can be shown that $((\hat{\boldsymbol{\beta}}^{[1]}, \hat{\boldsymbol{\beta}}^{[1]}) - (\boldsymbol{\beta}^{[1]}, \boldsymbol{\beta}^{[1]}))'$ has an asymptotic normal distribution with mean 0 and covariance matrix $[\boldsymbol{D}_{\boldsymbol{\beta}}{}^{'}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}{}^{-1}\boldsymbol{D}_{\boldsymbol{\beta}}]^{-1}[\boldsymbol{D}_{\boldsymbol{\beta}}{}^{'}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}{}^{-1}(\boldsymbol{Y}-\boldsymbol{\mu})(\boldsymbol{Y}-\boldsymbol{\mu})^{'}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}{}^{-1}\boldsymbol{D}_{\boldsymbol{\beta}}][\boldsymbol{D}_{\boldsymbol{\beta}}{}^{'}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}{}^{-1}\boldsymbol{D}_{\boldsymbol{\beta}}]^{-1}$ (refer to Sutradhar, 2003, Sutradhar et al., 2014, Mamode Khan et al., 2016 and Sunecher et al., 2017). The standard errors of $(\boldsymbol{\beta}^{[1]}, \boldsymbol{\beta}^{[2]})$ are obtained from the diagonals of the Hessian part in Equation (16).
As for the over-dispersion parameters, a second GQL is used where the score function is modified such that

$$\boldsymbol{D}_c^{*'}\boldsymbol{\Sigma}_c^{*-1}(\boldsymbol{Y}^2-\boldsymbol{\mu}^*)=0, \tag{17}$$

with $\boldsymbol{Y}^2=[\boldsymbol{Y}^{[1]^2}, \boldsymbol{Y}^{[2]^2}]_{(2T\times 1)}$, $\boldsymbol{\mu}^*=[E(\boldsymbol{Y}^{[1]^2}), E(\boldsymbol{Y}^{[2]^2})]_{(2T\times 1)}$ and $\frac{\partial \mu^{*[k]}}{\partial c_k}=\mu_t^{[k]^2}$.
For the diagonal entries in $\boldsymbol{\Sigma}_c^*$, they are derived using the MGF of the NB model which yields

$$\text{Var}(Y_t^{[k]^2})=\mu_t^{[k]}+(6+7c_k)\mu_t^{[k]^2}+(4+16c_k+12c_k^2)\mu_t^{[k]^3}+(4c_k+10c_k^2+6c_k^3)\mu_t^{[k]^4} \tag{18}$$

and as for the off-diagonal entries, they are computed using the 'working' multivariate normality assumption (refer to Prentice and Zhao, 1991 and Sunecher et al., 2017), where using

$$E[(Y_t-\mu_t)(Y_w-\mu_w)(Y_{w'}-\mu_{w'})]=0 \tag{19}$$

we obtain

$$E(Y_tY_wY_{w'})=E(Y_tY_w)\mu_{w'}+E(Y_tY_{w'})\mu_w+E(Y_wY_{w'})\mu_t-2\mu_t\mu_w\mu_{w'} \tag{20}$$

Hence,

$$\begin{aligned}\text{Cov}(Y_t^{[k]^2}, Y_{t+h}^{[k]^2}) &= (\mu_t^{[k]}+c_k\mu_t^{[k]^2})(\mu_{t+h}^{[k]}+c_k\mu_{t+h}^{[k]^2})+2[\rho_k^h(\mu_t^{[k]}+c_k\mu_t^{[k]^2})]^2 \\ &+\mu_{t+h}^{[k]^2}(\mu_t^{[k]}+c_k\mu_t^{[k]^2}+\mu_t^{[k]^2})+\mu_t^{[k]^2}(\mu_{t+h}^{[k]}+c_k\mu_{t+h}^{[k]^2}+\mu_{t+h}^{[k]^2}) \\ &+4\mu_t^{[k]}\mu_{t+h}^{[k]}[\rho_k^h(\mu_t^{[k]}+c_k\mu_t^{[k]^2})+\mu_t^{[k]}\mu_{t+h}^{[k]}]-5\mu_t^{[k]^2}\mu_{t+h}^{[k]^2} \\ &-(\mu_t^{[k]}+c_k\mu_t^{[k]^2}+\mu_t^{[k]^2})(\mu_{t+h}^{[k]}+c_k\mu_{t+h}^{[k]^2}+\mu_{t+h}^{[k]^2})\end{aligned} \tag{21}$$

and

$$\begin{aligned}\text{Cov}(Y_t^{[1]^2}, Y_t^{[2]^2}) &= (\mu_t^{[1]}+c_1\mu_t^{[1]^2})(\mu_t^{[2]}+c_2\mu_t^{[2]^2})+2[\text{Cov}(Y_t^{[1]}, Y_t^{[2]})]^2 \\ &+\mu_t^{[2]^2}(\mu_t^{[1]}+c_1\mu_t^{[1]^2}+\mu_t^{[1]^2})+\mu_t^{[1]^2}(\mu_t^{[2]}+c_2\mu_t^{[2]^2}+\mu_t^{[2]^2}) \\ &+4\mu_t^{[1]}\mu_t^{[2]}[\text{Cov}(Y_t^{[1]}, Y_t^{[2]})+\mu_t^{[1]}\mu_t^{[2]}]-5\mu_t^{[1]^2}\mu_t^{[2]^2} \\ &-(\mu_t^{[1]}+c_1\mu_t^{[1]^2}+\mu_t^{[1]^2})(\mu_t^{[2]}+c_2\mu_t^{[2]^2}+\mu_t^{[2]^2})\end{aligned} \tag{22}$$

The derivation of these formulae are shown in Sunecher et al. (2017). To facilitate the computation of Equation (22), the following boundary condition is assumed, $\text{Cov}(Y_t^{[1]}, Y_t^{[2]}) = \text{Cov}(Y_{t+1}^{[1]}, Y_{t+1}^{[2]})$ for $t = 0$ and, hence, $\text{Cov}(Y_t^{[1]}, Y_t^{[2]}) = \frac{\rho_{12,t}\sqrt{\lambda_t^{[1]}+c_1^*{\lambda_t^{[1]}}^2}\sqrt{\lambda_t^{[2]}+c_2^*{\lambda_t^{[2]}}^2}}{(1-\rho_1\rho_2)}$ for $t = 1$. From there, $\text{Cov}(Y_t^{[1]}, Y_t^{[2]})$ is computed iteratively.
The Newton-Raphson iteration to solve equation (17) is given by

$$\begin{pmatrix} \hat{c_1}_{r+1} \\ \hat{c_2}_{r+1} \end{pmatrix} = \begin{pmatrix} \hat{c_1}_r \\ \hat{c_2}_r \end{pmatrix} + [\boldsymbol{D}_c^{*'}\boldsymbol{\Sigma}_c^{*-1}\boldsymbol{D}_c^*]_r^{-1}[\boldsymbol{D}_c^{*'}\boldsymbol{\Sigma}_c^{*-1}(\boldsymbol{Y}^2 - \boldsymbol{\mu}^*)]_r \tag{23}$$

where $\hat{c_1}_r$ and $\hat{c_2}_r$ are the values at the r^{th} iteration. Under mild regularity and asymptotically conditions, these estimates are consistent and it can be shown that $((\hat{c}_1, \hat{c}_2) - (c_1, c_2))'$ has an asymptotic normal distribution with mean 0 and covariance matrix $[\boldsymbol{D}_c^{*'}\boldsymbol{\Sigma}_c^{*-1}\boldsymbol{D}_c^*]^{-1}[\boldsymbol{D}_c^{*'}\boldsymbol{\Sigma}_c^{*-1}(\boldsymbol{Y}^2 - \boldsymbol{\mu}^*)(\boldsymbol{Y}^2 - \boldsymbol{\mu}^*)'\boldsymbol{\Sigma}_c^{*-1}\boldsymbol{D}_c^*][\boldsymbol{D}_c^{*'}\boldsymbol{\Sigma}_c^{*-1}\boldsymbol{D}_c^*]^{-1}$ (see Sutradhar, 2003, Sutradhar et al., 2014, Mamode Khan et al., 2016 and Sunecher et al., 2017). The standard errors of (c_1, c_2) are obtained from the diagonals of the Hessian part in Equation (23).

4 Forecasting Equations

Using the BINAR(1) model in Equation (1) and (2),

$$Y_t^{[k]} = \alpha_k * Y_{t-1}^{[k]} + R_t^{[k]} \tag{24}$$

The conditional expectation and variance of the one step-ahead prediction of $Y_{t+1}^{[k]}$ given $Y_t^{[k]}$ is given by:

$$E(Y_{t+1}^{[k]}|Y_t^{[k]}) = \hat{\mu}_{t+1}^{[k]} + \hat{\rho}_k(Y_t^{[k]} - \hat{\mu}_t^{[k]}) \tag{25}$$

and

$$\begin{aligned}
\text{Var}(Y_{t+1}^{[k]}|Y_t^{[k]}) &= E_{\alpha_k}[\text{Var}(\alpha_k * Y_t^{[k]}|Y_t^{[k]}, \alpha_k)] + \text{Var}_{\alpha_k}[\text{E}(\alpha_k * Y_t^{[k]}|Y_t^{[k]}, \alpha_k)] + \text{Var}(R_{t+1}^{[k]}) \\
&= E_{\alpha_k}[\alpha_k(1-\alpha_k)Y_t^{[k]}] + \text{Var}_{\alpha_k}[\alpha_k Y_t^{[k]}] + \text{Var}(R_{t+1}^{[k]}) \\
&= [\hat{\rho}_k - \frac{\hat{\rho}_k(1-\hat{\rho}_k)\hat{c}_k}{1+\hat{c}_k} - \hat{\rho}_k^2]Y_t^{[k]} + \frac{\hat{\rho}_k(1-\hat{\rho}_k)\hat{c}_k}{1+\hat{c}_k}{Y_t^{[k]}}^2 \\
&+ [\hat{\mu}_{t+1}^{[k]} - \hat{\rho}_k\hat{\mu}_t^{[k]} + \hat{c}_k((\hat{\mu}_{t+1}^{[k]})^2 - \hat{\rho}_k(\hat{\mu}_t^{[k]})^2)] \\
&= \frac{\hat{\rho}_k(1-\hat{\rho}_k)}{1+\hat{c}_k}Y_t^{[k]}(1+\hat{c}_kY_t^{[k]}) + [\hat{\mu}_{t+1}^{[k]} - \hat{\rho}_k\hat{\mu}_t^{[k]} + \hat{c}_k((\hat{\mu}_{t+1}^{[k]})^2 - \hat{\rho}_k(\hat{\mu}_t^{[k]})^2)]
\end{aligned} \tag{26}$$

5 Case Study

One of the oldest and most popular football leagues in Europe is the English Premier League comprising of 20 teams. Arsenal Football Club (AFC) is the only football team

in the English Premier League which has remained among the top four during the last 20 years but in the recent Premier League seasons, the club's performance has been quite fluctuating. Several factors may explain this phenomena such as the home (coded as 1) or away (coded as 0) effect (x_{t1}), the number of inter-matches played (UEFA Champions League, League Cup and Football Association Cup) between the two Premier League consecutive matches (x_{t2}), the number of injured players in inter-matches other than Premier League matches (x_{t3}), the number of new players bought (x_{t4}) and the number of players retired or sold or loaned to other clubs (x_{t5}). To analyze the effect of these factors on the first ($Y_t^{[1]}$) and second half number of goals scored ($Y_t^{[2]}$), data from 2010-2011 to mid 2016-2017 seasons were collected, making a total of 240 paired observations. The table below provides a summary statistics, time series plots and autocorrelation functions (ACFs) of the 240 bivariate time series of counts of the number of goals scored in first and second half by AFC:

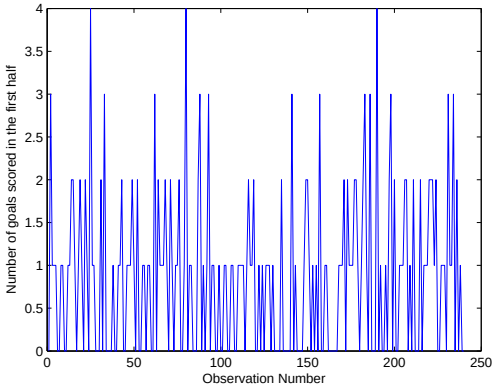

Figure 1: Time series plot for the number of goals scored by Arsenal in first half

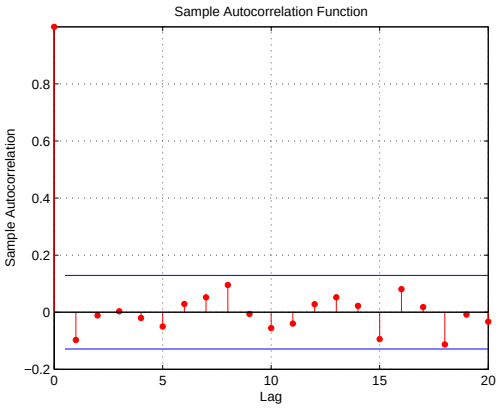

Figure 2: ACF plot

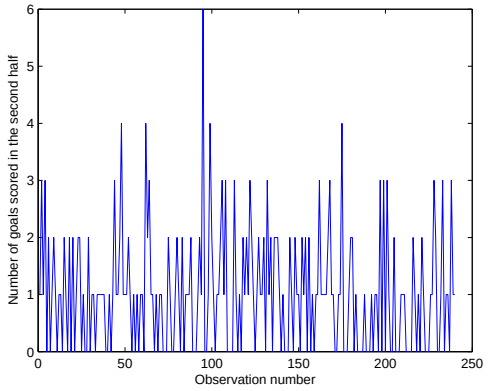

Figure 3: Time series plot for the number of goals scored by Arsenal in second half

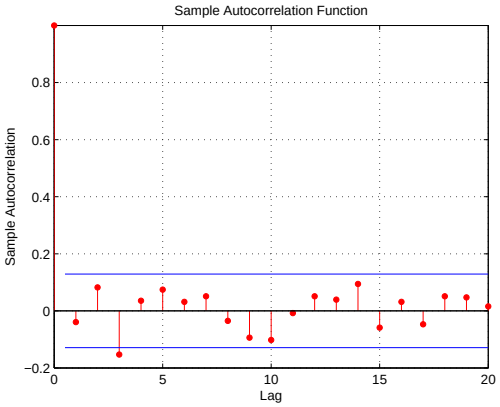

Figure 4: ACF plot

Table 1: Summary statistics for the number of goals scored in the first and second half for AFC.

	Sample Mean	Sample Variance	Lag-1	Cross
$Y_t^{[1]}$	0.8500	0.8644	0.0978	0.0063
$Y_t^{[2]}$	1.0125	1.0668	0.0394	

From the above summary statistics, it is clear that the variances of the two series are slightly greater than their respective means, confirming that the data is slightly over-dispersed. Hence, we use the BINAR(1) model with NB innovations explained in Section 2 together with the three GQL equations in Section 3 to analyze the over-dispersed bivariate data with the time-variant explanatory variables. The regression effects, over-dispersion, serial-correlation of the series and the cross-correlation of the innovations are displayed in the table below:

Table 2: Number of first and second half goals: GQL Estimates of the regression, over-dispersion, serial- and cross-correlation coefficients.

	Intercept	x_{t1}	x_{t2}	x_{t3}	x_{t4}	x_{t5}	c	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_{12,1}$
$Y_t^{[1]}$	0.273	0.213	-0.171	-0.061	0.091	-0.066	0.181	0.2741	0.3105	0.0077
$\exp(\hat{\beta})$	1.31	1.23	0.84	0.94	1.09	0.93				
s.e	(0.178)	(0.088)	(0.081)	(0.097)	(0.079)	(0.104)	(0.323)			
p-values	0.214	0.0157	0.0145	0.0056	0.0070	0.0015	0.0235			
$Y_t^{[2]}$	0.310	0.251	-0.150	-0.078	0.103	-0.075	0.191	0.2741	0.3105	0.0077
$\exp(\hat{\beta})$	1.36	1.28	0.86	0.92	1.10	0.92				
s.e	(0.191)	(0.073)	(0.059)	(0.083)	(0.091)	(0.090)	(0.308)			
p-values	0.219	0.0122	0.0136	0.0050	0.0034	0.0075	0.0287			

It is noted that all the parameters are significant, as shown in the above table. As

for the first covariate, in the first half AFC has 23 percent higher chance to score in a home match in comparison to an away match, and even a 28 percent higher chance in the second half for home compared to away matches. This is evident as we have seen in some matches AFC is losing in the first half, but then succeeded in winning the match in the second half. One such example is the match between Arsenal and Leicester in 2015-2016 season. In addition, the number of inter-matches played cause a decrease of 15 percent in the number of goals scored in first half and 13 percent in second half. AFC participates in the UEFA Champions League, Football Association Cup and League Cup every seasons and this has a negative impact on the performance of the team in Premier League matches. Next, the number of injured players have always been a major concern for AFC. Thus, when players are injured, this cause a decrease of 5 percent in the number of goals scored in the first half and 7 percent in the second half. Hence, one solution is to recruit new players. However, recruitment of players can be conducted only in August and January of each season. The impact of new players on the number of goals scored is 9 percent in first half and 10 percent in second half. Another variable which has a negative impact on the scoring capability of AFC is the number of players leaving the club. Whenever players retire, the chance of scoring goals decreases by 6 percent in first half and 7 percent in second half. Hence, AFC should be more active in the transfer market as this can have a big influence in its scoring capability. The over-dispersion parameters are all significant.

6 Conclusion

The paper reviews the non-stationary BINAR(1) process to model the first and second half number of football goals. The main contribution lies in the estimation of the regression and over-dispersion coefficients of each series using a two phase GQL approach. In particular, the auto-covariance structure that relates to the estimation of the over-dispersion parameters is constructed using the 'working' multivariate normality assumption since expressions for the high-ordered moments are not readily available in multivariate discrete set-ups. The model was applied to the Arsenal Football data where valid conclusions could be made regarding the current situation of the club in the Premier League.

References

Aly, E. and Bouzar, N. (2005). Stationary solutions for integer-valued autoregressive processes. *International Journal of Mathematics and Mathematical Sciences*, 20(1):1–18.

Baio, G. and Blangiardo, M. (2010). Bayesian hierarchical model for the prediction of football results. *Journal of Applied Statistics*, 37(2):253–264.

Baxter, M. and Stevenson, R. (1998). Discriminating between Poisson and negative binomial distributions: an application to goal scoring in association football. *Journal of Applied Statistics*, 15:347–438.

Bourguignon, M. and Vasconcellos, K. (2015). Improved estimation for Poisson INAR(1) models. *Journal of Statistical Computation and Simulation*, 85(12):2425–2441.

Groll, A. and Abedieh, J. (2013). Spain retains its title and sets a new record-generalized linear mixed models on European football championships. *Journal of Quantitative Analysis in Sports*, 9(1):51–66.

Groll, A., Schauberger, G., and Tutz, G. (2015). Brazil or Germany who will win the trophy? prediction of the FIFA World Cup 2014 based on team-specific regularized Poisson regression. *Journal of Quantitative Analysis in Sports*, 11(2):51–66.

Karlis, D. and Ntzoufras, I. (2000). On modelling soccer data. *Student*, 3:229–244.

Karlis, D. and Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society*, 52:381–393.

Karlis, D. and Ntzoufras, I. (2010). Robust fitting of football prediction models. *IMA Journal of Management Mathematics*, 22:171–182.

Louzada, F., Suzuki, A., and Salasar, L. (2014). Predicting match outcomes in the English Premier League: Which will be the final rank? *Journal of Data Science*, 12:235–254.

Maher, M. (1982). Modelling association football scores. *Statistica Neerlandica*, 36:109–118.

Mamode Khan, N., Sunecher, Y., and Jowaheer, V. (2016). Modelling a non-stationary BINAR(1) Poisson process. *Journal of Statistical Computation and Simulation*, 86:3106–3126.

McKenzie, E. (1986). Autoregressive moving-average processes with Negative Binomial and geometric marginal distrbutions. *Advanced Applied Probability*, 18:679–705.

McKenzie, E. (1988). Some ARMA models for dependent sequences of Poisson counts. *Advances in Applied Probability*, 20:822–835.

Pedeli, X. and Karlis, D. (2009). Bivariate INAR(1) models. Technical report, Athens University of Economics.

Pedeli, X. and Karlis, D. (2011). A bivariate INAR(1) process with application. *Statistical Modelling: An International Journal*, 11:325–349.

Prentice, R. and Zhao, L. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics*, 47:825–39.

Silva, M. and Oliveira, V. (2004). Difference equations for the higher order moments and cumulants of the INAR(1) model. *Journal of Time Series Analysis*, 25:317–333.

Steutel, F. and Van Harn, K. (1979). Discrete analogues of self-decomposability and statibility. *The Annals of Probability*, 7:3893–899.

Sunecher, Y., Mamodekhan, N., and Jowaheer, V. (2017). A gql estimation approach for analysing non-stationary over-dispersed BINAR(1) time series. *Journal of Statistical Computation and Simulation*.

Sutradhar, B. (2003). An overview on regression models for discrete longitudinal responses. *Statistical Science*, 18(3):377–393.

Sutradhar, B., Jowaheer, V., and Rao, P. (2014). Remarks on asymptotic efficient

estimation for regression effects in stationary and non-stationary models for panel count data. *Brazilian Journal of Probability and Statistics*, 28(2):241–254.

Weiβ, C. (2008b). Thinning operations for modelling time series of counts-a survey. *AStA Advances in Statistical Analysis*, 92(3):319–341.