The Statistician (2003)
52, *Part* 3, *pp.* 381–393

Analysis of sports data by using bivariate Poisson models

Dimitris Karlis

Athens University of Economics and Business, Greece

and Ioannis Ntzoufras

University of the Aegean, Chios, Greece

[Received November 2001. Final revision April 2003]

Summary. Models based on the bivariate Poisson distribution are used for modelling sports data. Independent Poisson distributions are usually adopted to model the number of goals of two competing teams. We replace the independence assumption by considering a bivariate Poisson model and its extensions. The models proposed allow for correlation between the two scores, which is a plausible assumption in sports with two opposing teams competing against each other. The effect of introducing even slight correlation is discussed. Using just a bivariate Poisson distribution can improve model fit and prediction of the number of draws in football games. The model is extended by considering an inflation factor for diagonal terms in the bivariate joint distribution. This inflation improves in precision the estimation of draws and, at the same time, allows for overdispersed, relative to the simple Poisson distribution, marginal distributions. The properties of the models proposed as well as interpretation and estimation procedures are provided. An illustration of the models is presented by using data sets from football and water-polo.

Keywords: Bivariate Poisson regression; Difference of Poisson variates; Inflated distributions; Soccer

1. Introduction

The Poisson distribution has been widely accepted as a simple modelling approach for the distribution of the number of goals in sports involving two competing teams. Although several researchers (see, for example, Lee (1997) and Karlis and Ntzoufras (2000) and the references therein) have shown the existence of a (relatively low) correlation between the number of goals scored by the two opponents, this has been ignored in most modelling approaches since it demands more sophisticated techniques. Maher (1982) discussed this issue, and Dixon and Coles (1997) extended the independent Poisson model by introducing indirectly a type of dependence. In team sports, such as football and water-polo, it is reasonable to assume that the two outcome variables are correlated since the two teams interact during the game. Moreover, in some sports games, the two opponents try to score sequentially, and thus the speed of the game of one team leads to more opportunities for both teams to score. A typical example is basketball: the correlations for the National Basketball Association and the Euroleague scores for the 2000–2001 season are 0.41 and 0.38 respectively.

Address for correspondence: Dimitris Karlis, Department of Statistics, Athens University of Economics and Business, Patission 76, 104 34 Athens, Greece.
E-mail: karlis@hermes.aueb.gr

0039–0526/03/52381

An alternative to the independent Poisson model can be constructed assuming that the two outcome variables follow a bivariate Poisson distribution (see Kocherlakota and Kocherlakota (1992) and references therein). The marginal distributions are simple Poisson distributions, whereas the random variables are now dependent. Maher (1982) mentioned the bivariate Poisson distribution but its use has been largely ignored, mainly because of the computational burden for fitting such a model.

The remainder of the paper proceeds as follows. Firstly, in Section 2, we present briefly the bivariate Poisson distribution and discuss its applicability in modelling sports data, especially for football games. The bivariate Poisson distribution allows for improving the model fit of the number of draws, a problem reported by some researchers (for example see Maher (1982) and Lee (1997)). An interesting feature of the bivariate Poisson model is the fact that the distribution of the difference of the two variates is the same as the distribution of the difference of two independent Poisson variates. However, the parameters have an entirely different interpretation. Moreover, an incorrect use of the independent Poisson case leads to significant differences. The effect of such a misspecification is illustrated by using a simple example. Maximum likelihood estimation of the parameters is made through an EM algorithm. In Section 3, extensions through inflated models are proposed. Since a draw is represented by diagonal terms in a bivariate distribution, adding an inflation term on the diagonal allows for more precise modelling of the number of draws. In Section 4, the models proposed are illustrated by using examples from football and water-polo. Finally, concluding remarks can be found in Section 5.

2. The bivariate Poisson distribution and its implementation in sports modelling

2.1. The bivariate Poisson distribution

Consider random variables X_κ, $\kappa = 1, 2, 3$, which follow independent Poisson distributions with parameters $\lambda_\kappa > 0$. Then the random variables $X = X_1 + X_3$ and $Y = X_2 + X_3$ follow jointly a bivariate Poisson distribution $\text{BP}(\lambda_1, \lambda_2, \lambda_3)$, with joint probability function

$$\begin{aligned} P_{X,Y}(x, y) &= P(X = x, Y = y) \\ &= \exp\{-(\lambda_1 + \lambda_2 + \lambda_3)\} \frac{\lambda_1^x}{x!} \frac{\lambda_2^y}{y!} \sum_{k=0}^{\min(x, y)} \binom{x}{k} \binom{y}{k} k! \left(\frac{\lambda_3}{\lambda_1 \lambda_2} \right)^k. \end{aligned} \qquad (1)$$

This bivariate distribution allows for dependence between the two random variables. Marginally each random variable follows a Poisson distribution with $E(X) = \lambda_1 + \lambda_3$ and $E(Y) = \lambda_2 + \lambda_3$. Moreover, $\text{cov}(X, Y) = \lambda_3$, and hence λ_3 is a measure of dependence between the two random variables. If $\lambda_3 = 0$ then the two variables are independent and the bivariate Poisson distribution reduces to the product of two independent Poisson distributions (referred to as the double-Poisson distribution). For a comprehensive treatment of the bivariate Poisson distribution and its multivariate extensions see Kocherlakota and Kocherlakota (1992) and Johnson *et al.* (1997).

It is plausible to adopt this distribution for modelling dependence in team sports. A natural interpretation of the parameters of a bivariate Poisson model is that λ_1 and λ_2 reflect the 'net' scoring ability of each team whereas λ_3 reflects game conditions (e.g. the speed of the game, the weather or stadium conditions).

2.2. The probability of the difference

Let us now define the difference $Z = X - Y$ of the goals scored by two opposing teams. Since $P(Z = z) = P(X - Y = z) = P(X_1 + X_3 - X_2 - X_3 = z) = P(X_1 - X_2 = z)$, the probability function of Z is independent of λ_3 and is the same as that derived from two independent Poisson variates. So Z follows the Poisson difference distribution with parameters λ_1 and λ_2, denoted as $\mathrm{PD}(\lambda_1, \lambda_2)$, given by

$$P_Z(z) = P(Z = z) = \exp\{-(\lambda_1 + \lambda_2)\} \left(\frac{\lambda_1}{\lambda_2}\right)^{z/2} I_z\left\{2\surd(\lambda_1\lambda_2)\right\}, \tag{2}$$

$z = \ldots, -3, -2, -1, 0, 1, 2, 3, \ldots$, where $I_r(x)$ denotes the modified Bessel function (see Abramowitz and Stegun (1974), page 375) defined by

$$I_r(x) = \left(\frac{x}{2}\right)^r \sum_{k=0}^{\infty} \frac{(x^2/4)^k}{k!\,\Gamma(r + k + 1)}. \tag{3}$$

A special case of two independent Poisson distributions for the case of equal means was derived by Irwin (1937) whereas the general case was considered by Skellam (1946). Interesting references on the Poisson difference distribution can be found in Johnson *et al.* (1992), page 191. Keller (1994) calculated the probability of winning a game assuming independent Poisson distributions for both X and Y.

Although distribution (2) implies that the winning probability ($Z > 0$) does not depend on parameter λ_3, treating the number of goals independently for each team leads to an overestimation of model parameters. It should be kept in mind that, since the parameters λ_1 and λ_2 are estimated from the marginal distributions, the covariance parameter λ_3 is confounded. In the following section we examine the effect of such a misspecification.

2.3. The effect of model misspecification

Let us consider that the true underlying model is the bivariate Poisson model but we use instead the double-Poisson model. Then we assume that the difference $Z = X - Y \sim \mathrm{PD}(\lambda_1 + \lambda_3, \lambda_2 + \lambda_3)$ instead of the correct $Z \sim \mathrm{PD}(\lambda_1, \lambda_2)$. This misspecification has quite a large effect even if the covariance λ_3 is as low as 0.10, which is about the observed covariance in football.

Fig. 1 depicts the relative change in the probability of a draw between the two competing teams, when independent Poisson distributions are considered compared with the bivariate

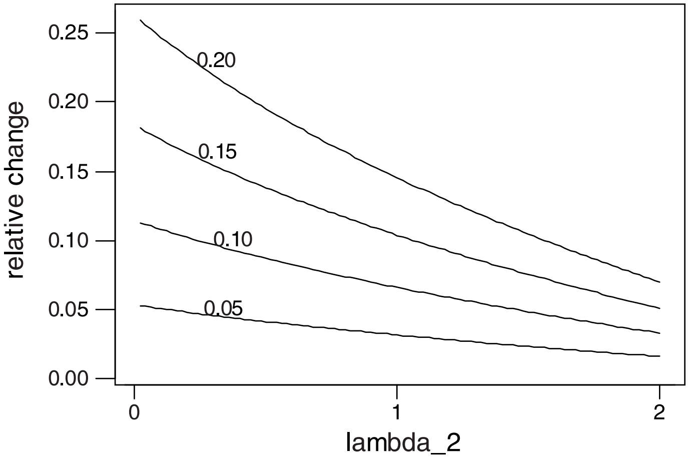

Fig. 1. Relative change in the probability of a draw when the two competing teams have marginal means equal to $\lambda_1 = 1$ and λ_2 ranging from 0.1 to 2: the different lines correspond to different levels of the covariance λ_3

Poisson model with small covariance parameter λ_3. Namely the independent Poisson model assumes that $X \sim \text{Poisson}(1)$ whereas $Y \sim \text{Poisson}(\lambda_2)$. The competing model assumes that $(X, Y) \sim \text{BP}(1 - \lambda_3, \lambda_2 - \lambda_3, \lambda_3)$. Note that for both models the marginal means are the same, but the bivariate Poisson model assumes also the existence of the covariance λ_3. We let the value of λ_2 vary from 0.1 to 2, for different values of the covariance term $\lambda_3 = 0.05, 0.10, 0.15, 0.20$. These values of the covariance are rather small but are close to those that are observed in real football data.

Fig. 1 shows that, for the typically observed range of counts in football data, the probability of a draw under a bivariate Poisson model is larger than the corresponding probability under the double-Poisson model even if λ_3 is quite small. For example, if we consider a bivariate Poisson model with $\lambda_3 = 0.05$ and $\lambda_1 = \lambda_2 = 1$ then we expect almost 3.3% more draws than under the corresponding independent Poisson model, whereas if λ_3 increases to 0.20 we expect 14% more draws. It is also clear that the larger the λ_3 the larger the relative change is. This may explain the empirical fact that the observed number of draws is usually larger than that predicted under an independent Poisson model.

2.4. Estimation

In this section we focus on the estimation of the parameters of the bivariate Poisson distribution. For the simple, but unrealistic for sports data, model without covariates, standard estimation procedures have been proposed (see, for example, Kocherlakota and Kocherlakota (1992)). Here we consider more realistic models that include covariates. Bivariate Poisson regression models have been described recently in Kocherlakota and Kocherlakota (2001) and Ho and Singer (2001). The former presents a Newton–Raphson approach for maximizing the likelihood whereas the latter describes a generalized least squares method.

Let us consider the general case of a bivariate Poisson regression. For the ith observation the model takes the form

$$\begin{aligned} (X_i, Y_i) &\sim \text{BP}(\lambda_{1i}, \lambda_{2i}, \lambda_{3i}), \\ \log(\lambda_{\kappa i}) &= \mathbf{w}_{\kappa i} \boldsymbol{\beta}_\kappa, \qquad \kappa = 1, 2, 3, \end{aligned} \tag{4}$$

$i = 1, \ldots, n$ denotes the observation number, $\mathbf{w}_{\kappa i}$ denotes a vector of explanatory variables for the ith observation used to model $\lambda_{\kappa i}$ and $\boldsymbol{\beta}_\kappa$ denotes the corresponding vector of regression coefficients. It is clear that each parameter of the bivariate Poisson distribution may be influenced by different characteristics and variables. For this reason, the explanatory variables that are used to model each parameter $\lambda_{\kappa i}$ may not be the same. Parameter estimation for such a model is not straightforward. Hence, we make use of the EM algorithm to obtain maximum likelihood estimates.

To construct the EM algorithm for the bivariate Poisson regression model, we make use of the trivariate reduction derivation of the bivariate Poisson distribution. For this reason, for each observation i, we further introduce the latent variables X_{1i}, X_{2i} and X_{3i} for which we assume a Poisson distribution with parameters λ_{1i}, λ_{2i} and λ_{3i} respectively. Moreover, we assume that $X_i = X_{1i} + X_{3i}$ and $Y_i = X_{2i} + X_{3i}$.

The EM algorithm proceeds by estimating the unobserved data via their conditional expectations at the E-step and then it maximizes the complete-data likelihood at the M-step. Hence, at the E-step, we obtain the posterior expectation of X_{1i}, X_{2i} and X_{3i} given the data and the current parameter values and then, at the M-step, we maximize the complete-data likelihood by fitting three Poisson regression models. The aim is now to estimate the regression coefficients $\boldsymbol{\beta}_\kappa$

for $\kappa = 1, 2, 3$. The full algorithm is available from the authors on request. This EM algorithm is very flexible and many variations of the bivariate Poisson model can be fitted with slight modifications.

3. Inflated bivariate Poisson distributions

The bivariate Poisson model introduces correlation between the variables, but the marginal distributions are still Poisson. As an improvement, relative to the simple bivariate Poisson model, we may consider mixtures of bivariate Poisson distributions, either finite or infinite. Such mixtures may have a variety of forms, depending on the varying parameters and the mixing distribution. However, such models incorporate very complicated structures and, therefore, are not very useful for practical application in sports modelling.

A type of inflated model was used by Dixon and Coles (1997) for modelling football games. They initially assumed two independent Poisson distributions and then they corrected the expected values for the outcomes (0–0, 1–0, 0–1, 1–1) by an additional parameter. We propose alternative models which, at the same time, incorporate correlation between the variables and overdispersed (relative to Poisson) marginal distributions while they further improve the fit on the counts of draws.

If the score 0–0 is underestimated by the model then we may inflate the probability at the (0,0) cell by adding a parameter. In such a case, a model similar to that proposed by Li *et al.* (1999) is specified. We propose a more general model formulation which inflates the probabilities of draws. A draw between two teams is represented by the outcomes on the diagonal of the probability table. To correct for the excess of draws we may add an inflation component on the diagonal of the probability function. This model is an extension of the simple zero-inflated model that allows only for an excess in (0,0) draws. We consider, for generality, that the starting model is the bivariate Poisson model.

Under this approach a diagonal inflated model is specified by

$$P_D(x, y) = \begin{cases} (1-p)\ \mathrm{BP}(x, y|\lambda_1, \lambda_2, \lambda_3), & x \neq y, \\ (1-p)\ \mathrm{BP}(x, y|\lambda_1, \lambda_2, \lambda_3) + p\,D(x, \boldsymbol{\theta}), & x = y, \end{cases} \tag{5}$$

where $D(x, \boldsymbol{\theta})$ is a discrete distribution with parameter vector $\boldsymbol{\theta}$. Such models can be fitted by using the EM algorithm.

Useful choices for $D(x, \boldsymbol{\theta})$ are the Poisson, the geometric or simple discrete distributions such as the Bernoulli distribution. The geometric distribution might be of great interest since it has its mode at zero and decays quickly. As a discrete distribution we consider $P(X = j) = \theta_j$ for $j = 0, 1, 2, \ldots, J$, where $\Sigma_{j=0}^{J} \theta_j = 1$; $J \leqslant 3$ is usually sufficient for football data whereas $J = 0$ corresponds to zero-inflated models. Although univariate zero-inflated Poisson regression models have been developed and examined in detail (see, for example, Lambert (1992) and Böhning *et al.* (1999)), multivariate extensions, similar to the models proposed in this paper, are relatively rare with the exception of Li *et al.* (1999), Gan (2000) and Wahlin (2001).

There are two important properties of such models. Firstly, the marginal distributions of a diagonal inflated model are not Poisson distributions but mixtures of distributions with one Poisson component. Secondly, if $\lambda_3 = 0$ (corresponding to the double-Poisson distribution) the resulting inflated distribution introduces a degree of dependence between the two variables under consideration. For this reason, diagonal inflation may correct both the overdispersion and the correlation problems that are encountered in modelling football games.

4. Application in sports

4.1. Modelling football games

Many references in this field assume that the number of goals scored by each team follows a Poisson distribution (Maher (1982), Lee (1997) and Rue and Salvesen (2000) among others). Such models take the general form

$$X_i \sim \text{Poisson}(\lambda_{1i}),$$

$$Y_i \sim \text{Poisson}(\lambda_{2i}),$$

$$\log(\lambda_{1i}) = \mu + \text{home} + \text{att}_{h_i} + \text{def}_{g_i} + \text{home.att}_{h_i} + \text{home.def}_{g_i} + \text{att.def}_{h_i g_i} + \text{home.att.def}_{h_i g_i}$$

$$\log(\lambda_{2i}) = \mu + \text{att}_{g_i} + \text{def}_{h_i} + \text{att.def}_{g_i h_i}$$

for $i = 1, 2, \ldots, n$, where n is the number of games or observations, i is a game (observation) indicator, h_i and g_i indicate the home and the away team in game i, X_i and Y_i are the goals scored by the home (h_i) and the away (g_i) team in game i, λ_{1i} and λ_{2i} are the corresponding expected number of goals, μ is a constant parameter, home is the home effect parameter and finally att_k and def_k encapsulate the offensive (or attacking) and defensive performances of team k. Karlis and Ntzoufras (2000) examined such models in a general log-linear setting allowing also for model selection.

Although we have implemented our proposed models in various data sets, here we focus on the Italian serie A data for the 1991–1992 season and give some brief details for the Champions League data for the 2000–2001 season. The classical likelihood ratio test (LRT) and its asymptotic χ^2 p-value as well as Bayes information criterion (BIC) and Akaike information criterion (AIC) were used in the selection and fitting of models. We adopted a simpler structure for the parameters involved in the linear predictors of λ_1 and λ_2. Hence, for each game i ($i = 1, \ldots, n$),

$$\begin{aligned} (X_i, Y_i) &\sim \text{BP}(\lambda_{1i}, \lambda_{2i}, \lambda_{3i}), \\ \log(\lambda_{1i}) &= \mu + \text{home} + \text{att}_{h_i} + \text{def}_{g_i}, \\ \log(\lambda_{2i}) &= \mu + \text{att}_{g_i} + \text{def}_{h_i}. \end{aligned} \tag{6}$$

To achieve identifiability of the above model parameters, we may use any standard set of constraints. Here we propose to use either sum-to-zero or corner constraints, depending on the interpretation that we prefer. For this example, we chose sum-to-zero constraints for ease of interpretation. Therefore, the overall constant parameter specifies λ_1 and λ_2 when two teams of the same strength play on a neutral field. Offensive and defensive parameters are expressed as departures from a team of average offensive or defensive ability.

For the covariance parameters λ_{3i} we considered various versions of the linear predictor which can be summarized by

$$\log(\lambda_{3i}) = \beta^{\text{con}} + \gamma_1 \beta^{\text{home}}_{h_i} + \gamma_2 \beta^{\text{away}}_{g_i},$$

where β^{con} is a constant parameter and $\beta^{\text{home}}_{g_i}$ and $\beta^{\text{away}}_{h_i}$ are parameters that depend on the home and away team respectively. Parameters γ_1 and γ_2 are dummy binary indicators taking values 0 or 1, depending on the model that we consider. Hence when $\gamma_1 = \gamma_2 = 0$ we consider constant covariance, when $(\gamma_1, \gamma_2) = (1, 0)$ we assume that the covariance depends on the home team only and so on.

The parameter λ_3 can be interpreted as a random effect which acts additively on the marginal mean and reflects game conditions. An alternative structure on the design matrix can be easily implemented if additional information is available, or if we assume that attacking abilities are different in home and away games, or if the home effect varies from team to team.

Diagonal inflated models are suitable for championships with an excess of draws which cannot be captured by double-Poisson models, or even bivariate Poisson models. Here we illustrate diagonal inflated models in the Italian serie A data for the 1991–1992 season. The scoring system in that season gave 2 points for a win and 1 point for a draw. This system often gives an excess of draws. We considered various models including the double-Poisson, the bivariate Poisson and the diagonal inflated models using several diagonal distributions; Table 1. The best-fitted model is the bivariate Poisson model with an extra parameter for the 1–1 score which was otherwise considerably underestimated: Table 2. The model selected is supported by the AIC, BIC and LRT for testing the hypothesis $H_0 : p = 0$, where p is the inflation proportion (p-value less than 0.01). Details for the model selection procedure are given in Table 1. Note that the zero-inflated and the geometric diagonally inflated models did not improve the likelihood since 0–0 scores were not underestimated. Moreover, the improvement that is offered by a Poisson diagonal component was statistically significant when added to the simple Poisson model but not when added to the bivariate Poisson model.

Table 1. Details of the fitted models for the Italian serie A 1991–1992 data

Model distribution	*Additional model details*	*Log-likelihood*	*Number of parameters*	*p-value*	*AIC*	*BIC*
1, double Poisson		−771.5	36		1614.9	1774.2
Covariates on λ_3						
2, bivariate Poisson	Constant ($\gamma_1 = \gamma_2 = 0$)	−764.9	37	0.00†	1603.9	1767.5
3, bivariate Poisson	Home team effect ($\gamma_1 = 1, \gamma_2 = 0$)	−758.9	55	0.84‡	1627.8	1871.1
4, bivariate Poisson	Away team effect ($\gamma_1 = 0, \gamma_2 = 1$)	−755.6	55	0.41‡	1621.2	1864.5
5, bivariate Poisson	Home and away team effects ($\gamma_1 = \gamma_2 = 1$)	−745.9	72	0.33‡	1635.7	1954.3
6, zero-inflated bivariate Poisson	Constant	−764.9	38	1.00§	1605.9	1773.9
Diagonal distribution						
7, diagonal inflated bivariate Poisson	Geometric	−764.9	39	1.00§	1607.9	1780.3
8, diagonal inflated bivariate Poisson§§	Discrete (1)	−756.6	39	0.00§	1591.1	1763.7
9, diagonal inflated bivariate Poisson	Discrete (2)	−756.6	40	1.00*	1593.1	1770.1
10, diagonal inflated bivariate Poisson	Discrete (3)	−756.4	41	0.54**	1594.8	1776.2
11, diagonal inflated bivariate Poisson	Poisson	−763.5	39	0.25§	1605.1	1777.5
12, diagonal inflated Poisson	Poisson	−767.0	38	0.01§	1610.0	1778.1

†$H_0 : \lambda_3 = 0$.
‡$H_0 : \lambda_3 =$ constant.
§$H_0 : p = 0$.
§§Best-fitted model.
*$H_0 : \theta_2 = 0$.
**$H_0 : \theta_3 = 0$.

Table 2. Estimated draws for every model

Model distribution	*Additional model details*	*Estimates for the following scores:*				
		0–0	*1–1*	*2–2*	*3–3*	*4–4*
Observed data		38	58	10	4	1
1, double Poisson		38	33	9	1	0
Covariates on λ_3						
2, bivariate Poisson	Constant	49	35	11	2	0
3, bivariate Poisson	Home team effect	51	34	11	3	0
4, bivariate Poisson	Away team effect	49	34	11	2	0
5, bivariate Poisson	Home and away team effects	47	32	10	2	0
6, zero-inflated bivariate Poisson		49	35	11	2	0
Diagonal distribution						
7, diagonal inflated bivariate Poisson	Geometric	49	35	11	2	0
8, diagonal inflated bivariate Poisson†	Discrete (1)	43	58	9	2	0
9, diagonal inflated bivariate Poisson	Discrete (2)	43	58	9	2	0
10, diagonal inflated bivariate Poisson	Discrete (3)	43	58	9	3	0
11, diagonal inflated bivariate Poisson	Poisson	50	38	13	3	1
12, diagonal inflated Poisson	Poisson	45	40	14	3	1

†Best-fitted model.

The goodness of fit was assessed by comparing our proposed model with the full or saturated model which fits the data exactly. According to the LRT our proposed model fits our data sufficiently well (p-value 0.85). Moreover, the AIC and BIC measures for the full model are 2204.0 and 4928.7 respectively. Both these criteria indicate the selection of our model against the alternative full or saturated model.

Table 3 provides the parameter estimates of a simple Poisson model and the selected diagonal inflated bivariate Poisson model. The expected number of goals from this model for game i with home team h_i and away team g_i are

$$\begin{aligned} E(X_i) &= \lambda_{1i}, \\ E(Y_i) &= \lambda_{2i}, \end{aligned} \tag{7}$$

where λ_{1i} and λ_{2i} are given by expression (6). For the inflated bivariate Poisson model, a Bernoulli distribution was used as inflation with parameter θ_1. The expected number of goals by using this model can be calculated as

$$\begin{aligned} E(X_i) &= (1-p)(\lambda_{1i}+\lambda_{3i})+p\theta_1, \\ E(Y_i) &= (1-p)(\lambda_{2i}+\lambda_{3i})+p\theta_1, \end{aligned} \tag{8}$$

where λ_{1i} and λ_{2i}

Table 3. Estimated parameters for the Poisson and bivariate Poisson models for 1991–1992 Italian serie A data†

Team	*Results for model 1, double Poisson*		*Results for model 2, bivariate Poisson*	
	Attack	*Defence*	*Attack*	*Defence*
1, Milan	0.68	−0.50	0.84	−1.18
2, Juventus	0.18	−0.50	0.22	−0.70
3, Torino	0.11	−0.60	0.18	−0.86
4, Napoli	0.43	0.12	0.51	0.19
5, Roma	0.00	−0.16	0.02	−0.17
6, Sampdoria	0.02	−0.16	0.10	−0.16
7, Parma	−0.15	−0.27	−0.14	−0.34
8, Inter	−0.29	−0.28	−0.37	−0.29
9, Foggia	0.49	0.50	0.57	0.63
10, Lazio	0.16	0.10	0.28	0.21
11, Atalanta	−0.18	−0.11	−0.21	−0.11
12, Fiorentina	0.18	0.13	0.29	0.28
13, Genoa	−0.04	0.25	−0.09	0.40
14, Cagliari	−0.21	−0.08	−0.21	−0.01
15, Verona	−0.40	0.43	−0.51	0.57
16, Bari	−0.33	0.24	−0.50	0.33
17, Cremonese	−0.29	0.28	−0.36	0.45
18, Ascoli	−0.34	0.61	−0.64	0.75
Other parameters				
Intercept μ	−0.18		−0.57	
Home team effect	0.36		0.50	
λ_3	0.00		0.23	
Mixing proportion	0.00		0.09	
θ_1			1.00	

†Expected number of goals can be calculated by using equations (7) and (6) for model 1 and equations (8) and (6) for model 2.

It is worth mentioning that the scoring system has been changed to encourage teams not to be satisfied by draws; a win is now worth 3 points and a draw just 1 point. This has led to a reduction in the number of draws in recent championships.

Note that the LRT for testing mixture models with different numbers of components is known to be inappropriate (see, for example, Lindsay (1995)). So the choice of model between such mixtures can be based on the AIC.

A variety of models was fitted in the Champions League data for the 2000–2001 season. The best-fitted model was the bivariate Poisson model with constant λ_3 supported by the LRT which rejects the hypothesis $H_0 : \lambda_3 = 0$ (p-value 0.042) and the AIC. The zero- and diagonal inflated models did not improve the model fit. This is mainly because these models are useful only when the model selected underestimates the number of draws.

4.2. Modelling water-polo outcomes

In this section we present an implementation of the bivariate Poisson models on water-polo games. This sport was selected because of the relatively small scores (so that it is plausible to use discrete distributions) and large correlations between the scores of the competing teams. Useful information can be found at `www.usawaterpolo.com` and `www.hickoksports.com`.

The main aim of the game is to score goals. Usual scores are around 8 goals for each team, with strong correlation between the scores of the competing teams.

Here we implemented the bivariate Poisson models to the data of the European national team cup held at Florence in September 1999. 12 national teams played a total of 50 games. In our analysis we considered only full-time scores, ignoring extra time. The effect of this truncation is minimal with only two draws observed.

The model that we consider was similar to the corresponding model for football games without the parameter estimating the home effect:

$$(X_i, Y_i) \sim \mathrm{BP}(\lambda_{1i}, \lambda_{2i}, \lambda_{3i}),$$

$$\log(\lambda_{1i}) = \mu + \mathrm{att}_{o_{1i}} + \mathrm{def}_{o_{2i}}, \tag{9}$$

$$\log(\lambda_{2i}) = \mu + \mathrm{att}_{o_{2i}} + \mathrm{def}_{o_{1i}}, \tag{10}$$

for $i = 1, \ldots, 50$; here o_{1i} and o_{2i} are indicators corresponding to the first and second mentioned team opponents in game i. For λ_3 we consider two cases; in the first λ_3 is constant whereas in the second λ_3 is given by

$$\log(\lambda_{3i}) = m + \mathrm{team}_{o_{1i}} + \mathrm{team}_{o_{2i}} \tag{11}$$

for each game between teams o_{1i} and o_{2i}, where team_k is the effect of team k on λ_3. For these data we did not consider the zero-inflated and diagonal inflated models since draws in water-polo are rare. As in the football implementation, we may use either sum-to-zero or corner constraints depending on the interpretation that we prefer. Here we considered corner constraints with Germany as the base-line team. The constant parameter specifies parameters λ_1 and λ_2 when Germany plays a team with identical offensive and defensive ability. Moreover, the offensive (att_k) and defensive (def_k) parameters express differences in each team's offensive or defensive abilities from those of Germany (the base-line team).

As a consequence three models were fitted (the double Poisson, bivariate Poisson with constant covariance and bivariate Poisson with covariance depending on opposing teams). All three criteria used (the LRT, AIC and BIC) indicated as best model the bivariate Poisson model with constant λ_3; Table 4. The AIC and BIC values for the full or saturated model (508.4 and 726.4 respectively) also indicated selection of our model against the full model. The covariance parameter λ_3 was found to be equal to 5.55, which indicated significant covariance between the scores of the opposing teams.

Table 4. Details of the fitted models for the water-polo 1999 European national cup

Model	*Covariates on λ_3*	*Log-likelihood*	*Number of parameters*	*p-value*	*AIC*	*BIC*
1, double Poisson		−186.67	23		417.3	474.3
2, bivariate Poisson	Constant	−171.91	24	0.000†	391.8	451.3
2, bivariate Poisson (Hungary base-line; $\mu = 0$)	Constant	−172.31	23	0.376‡	390.62	447.60
3, bivariate Poisson	Equation (11)	−167.31	35	0.602§	404.6	491.3

†$H_0 : \lambda_3 = 0$.
‡$H_0 : \mu = 0$.
§$H_0 : \lambda_3 =$ constant.

The parameter estimating the defensive ability of Hungary was found to be a very large negative number. This implies that the fitted values of the goals conceded by Hungary tend to be constant for all opposing teams and equal to the covariance parameter λ_3 (instead of $\lambda_2 + \lambda_3$). Any value for the defensive ability of Hungary lower than -20.0 results in identical fitted values and likelihood. For this reason, and to avoid unidentifiability, we have set the defensive ability of Hungary to -20.0. Although this may imply that the model does not provide a good or sensible estimation of the Hungarian team's defensive ability this is not so. If we isolate the data concerning the defence of the Hungarian team (7,3,6,3,6,4,5,12) we observe that they are underdispersed relative to the Poisson distribution (mean, 5.75; variance, 8.50). The variance test (see Karlis and Xekalaki (2000) for a critical review) does not reject the hypothesis of the Poisson distribution. Hence, the constant mean for the Hungarian team's defence seems plausible. For the rest of the teams, the observed data are overdispersed and therefore it is plausible to assume that the mean is not constant. Note that this numerical problem appears mainly because of the large number of parameters relative to the amount of data that is available in such competitions and to the unbalanced design of the data. For this reason, in full season leagues we shall not have such problems.

To avoid possible overparameterization we also considered a model with Hungary as a base-line team and the corresponding constant (which is a measure of the overall performance of Hungary) constrained to be 0. This may be interpreted in the following way: if Hungary plays with a team which has the same attacking and defensive ability then the expected number of goals scored will be 1 goal added to the covariance parameter λ_3, i.e. $E(X) = E(Y) = 1 + \lambda_3$. The advantage of such a model is that it does not assume a constant expected number of goals conceded by Hungary. Comparing the two models with a χ^2-test ($H_0 : \mu = 0$; p-value 0.38), the BIC or AIC we conclude that the model with Hungary as the base-line team and its

Table 5. Estimated parameters for the double-Poisson and bivariate Poisson models for the 1999 water-polo European national cup data†

Team	*Final ranking*	*Results for model 1, double Poisson*		*Results for model 2, bivariate Poisson ($\mu = 0$)*	
		Attack	*Defence*	*Attack*	*Defence*
1, Germany	8	−0.80	0.23	−2.03	2.28
2, Greece	4	−0.48	0.06	−0.79	1.81
3, Hungary	1	0.00	0.00	0.00	0.00
4, Italy	3	−0.45	−0.03	−1.68	0.12
5, Croatia	2	−0.15	0.18	−0.76	0.81
6, Netherlands	12	−0.56	0.50	−2.31	2.59
7, Romania	9	−0.74	0.16	−2.65	1.71
8, Russia	5	−0.34	0.39	−1.50	1.76
9, Slovakia	10	−0.54	0.29	−2.63	1.95
10, Slovenia	11	−0.63	0.34	−2.70	2.22
11, Spain	6	−0.28	0.35	−0.88	2.08
12, Yugoslavia	7	−0.54	0.17	−1.37	1.89
Other parameters					
Intercept		2.18		0.00	
λ_3		0.00		5.50	

†Base-line level Hungary. The expected numbers of goals are $\lambda_{1i} + \lambda_{3i}$ and $\lambda_{2i} + \lambda_{3i}$; λ_{1i} and λ_{2i} are given by equations (9) and (10) for both models. For model 1 $\lambda_{3i} = 0.0$.

Table 6. Estimated parameters for the double-Poisson and bivariate Poisson models for the 1999 water-polo European national cup data

Game	*Phase*	*Opponents*		*Observed score*	*Expected goals for the following models:*			
		Team 1	*Team 2*		*Double Poisson*		*Bivariate Poisson*†	
					Team 1	*Team 2*	*Team 1*	*Team 2*
1	Group A	Italy	Hungary	7–7	5.7	8.6	5.7	6.7
2	Group A	Slovenia	Hungary	3–11	4.7	12.4	5.6	14.8
3	Group A	Hungary	Croatia	7–6	10.6	7.7	7.8	6.0
4	Group A	Greece	Hungary	3–8	5.5	9.3	6.0	11.7
5	Group A	Hungary	Slovakia	13–6	11.8	5.2	12.6	5.6
6	Quarter-final	Hungary	Germany	15–4	11.1	4.0	15.4	5.7
7	Semi-final	Hungary	Italy	7–5	8.6	5.7	6.7	5.7
8	Final	Croatia	Hungary	12–15	7.7	10.6	6.0	7.8

†Hungary base-line; $\mu = 0$.

corresponding constant equal to 0 should be selected. For details of the model parameters of the two finally selected models see Table 5; fitted values for the games of Hungary are given in Table 6.

5. Discussion

In the present paper the bivariate Poisson distribution and its extensions were used to model sports data. The bivariate Poisson distribution allows for correlation between the scores of the competing teams, which is plausible for certain team sports. Diagonal inflated models are also proposed to improve the modelling aspects further. The models proposed provide a better fit of the football data since they can handle both correlation and overdispersion. Furthermore, they improve the fit on the diagonal of the observed table of results, which reflects ties between the two opponents. According to the models proposed we can extend the double-Poisson model by either considering a bivariate Poisson model or by inflating the diagonal elements of the joint probability function. Both models incorporate correlation, but the latter introduces overdispersion as well.

Maximum likelihood estimation for bivariate Poisson regression models was described in detail. The EM algorithm that is proposed in this paper can be easily extended to incorporate more complicated models. Such models were described in Kocherlakota and Kocherlakota (2001). Additionally, we have extended the zero-inflated multivariate Poisson models of Li *et al.* (1999) by defining more general inflation models, with potential implementation in manufacturing or marketing. The EM algorithm that was proposed for such models can be quite helpful for real data applications of such models.

Acknowledgements

The authors thank the Joint Editor and the referees for their helpful comments on previous versions of the paper.

References

Abramowitz, M. and Stegun, I. A. (1974) *Handbook of Mathematical Functions*. New York: Dover Publications.

Böhning, D., Dietz, E., Schlattmann, P., Mendonça, L. and Kirchner, U. (1999) The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. *J. R. Statist. Soc.* A, **162**, 195–209; correction, **163** (2000), 121–122.

Dixon, M. J. and Coles, S. G. (1997) Modelling association football scores and inefficiencies in the football betting market. *Appl. Statist.*, **46**, 265–280.

Gan, N. (2000) General zero-inflated models and their applications. *PhD Thesis*. North Carolina State University, Raleigh.

Ho, L. L. and Singer, J. M. (2001) Generalized least squares methods for bivariate Poisson regression. *Communs Statist. Theory Meth.*, **30**, 263–277.

Irwin, J. O. (1937) The frequency distribution of the difference between two independent variates following the same Poisson distribution. *J. R. Statist. Soc.* A, **100**, 415–416.

Johnson, N., Kotz, S. and Balakrishnan, N. (1997) *Discrete Multivariate Distributions*. New York: Wiley.

Johnson, N., Kotz, S. and Kemp, A. W. (1992) *Univariate Discrete Distributions*. New York: Wiley.

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

Karlis, D. and Xekalaki, E. (2000) A simulation comparison of several procedures for testing the Poisson assumption. *Statistician*, **49**, 355–382.

Keller, J. (1994) A characterization of the Poisson distribution and the probability of winning a game. *Am. Statistn*, **48**, 294–299.

Kocherlakota, S. and Kocherlakota, K. (1992) *Bivariate Discrete Distributions*. New York: Dekker.

Kocherlakota, S. and Kocherlakota, K. (2001) Regression in the bivariate Poisson distribution. *Communs Statist. Theory Meth.*, **30**, 815–827.

Lambert, D. (1992) Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1–14.

Lee, A. J. (1997) Modeling scores in the Premier League: is Manchester United really the best? *Chance*, **10**, 15–19.

Li, C. S., Lu, J. C., Park, J., Kim, K. and Peterson, J. (1999) Multivariate zero-inflated Poisson models and their applications. *Technometrics*, **41**, 29–38.

Lindsay, B. G. (1995) Mixture models: theory, geometry and applications. *Regl Conf. Ser. Probab. Statist.*, **5**.

Maher, M. J. (1982) Modelling association football scores. *Statist. Neerland.*, **36**, 109–118.

Rue, H. and Salvesen, Ø. (2000) Prediction and retrospective analysis of soccer matches in a league. *Statistician*, **49**, 399–418.

Skellam, J. G. (1946) The frequency distribution of the difference between two Poisson variates belonging to different populations. *J. R. Statist. Soc.* A, **109**, 296.

Wahlin, J. F. (2001) Bivariate ZIP models. *Biometr. J.*, **43**, 147–160.