Modelling soccer matches using bivariate discrete distributions with general dependence structure

Ian McHale and Phil Scarf

Centre for Operational Research and Applied Statistics,
Salford Business School,
University of Salford,
Salford,
Manchester
M5 4WT, UK.

email i.mchale@salford.ac.uk, p.a.scarf@salford.ac.uk

Salford Business School Working Paper Series

Paper no. 321/06

Modelling soccer matches using bivariate discrete distributions with general dependence structure

Ian McHale*, Phil Scarf
*Centre for Operational Research and Applied Statistics, The University of Salford, Greater Manchester, M5 4WT, UK. *i.mchale@salford.ac.uk*

In this paper copulas are used to generate novel bivariate discrete distributions. These distributions are fitted to soccer data from the English Premier League. An interesting aspect of these data is that the primary variable of interest, the discrete pair shots-for and shots-against, exhibit negative dependence; thus in particular we develop bivariate Poisson-related distributions that allow such dependence. The paper focuses on Archimedian copulas, for which the dependence structure is fully determined by a 1-dimensional projection that is invariant under marginal transformations. Diagnostic plots for copula fit based on this projection are adapted to deal with discrete variables. Covariates relating to within-match contributions such as numbers of passes and tackles are introduced to explain variability in shot outcomes. The results of this analysis would appear to support the notion that playing the "beautiful game" is an effective strategy—more passes and crosses contribute to more effective play and more shots on goal.

Keywords: soccer, copula, bivariate Poisson, negative dependence.

1. Introduction

Soccer is the most popular spectator sport in the world. From South America to Eastern Europe, Africa to Asia, fans are fanatical and television companies compete strongly to win the rights to broadcast games. Huge sums of money are involved, from players wages to transfer fees. And why does the game generate such interest? The simplicity of the objectives and rules, the uncertainty in match outcome - authors have proposed many explanations, see, for example, KONING (2000) and SZYMANSKI (2003). Ultimately, a team has to shoot at a goal in the hope the ball enters the goal. A shot can arise in many ways, but it is generally thought that passing the ball helps to create shooting opportunities. But what is the relationship between passes and shots? Further, what are the relationships with other variables, such as crosses, tackles and dribbles? In contrast to other team sports such as baseball, cricket and American Football, soccer does not generate a plethora of data relating to the individual components of play that make up a match. What data are available relate mainly to match results. Within game

data are scarce but we have obtained a comprehensive data set on within-game actions for matches in the English Premier League. Our aim here is to shed light on what exactly successful teams do.

As a consequence of data paucity, much of the literature on soccer has focused on forecasting match outcome. Two alternative approaches model results directly or indirectly. Direct models analyse actual match results, i.e. win, draw or loss, usually looked at from the perspective of the home team, whereas the indirect forecast generates probabilities for the number of goals scored by the home team and by the away team and probabilities of win, loss or draw are inferred from these. The direct approach has typically been implemented using an ordered probit model. This approach has proved popular amongst economists, see, for example, DOBSON and GODDARD (2001). The indirect approach models goals scored by each team by employing bivariate Poisson-related models (MAHER, 1982; DIXON and COLES, 1997; KARLIS and NTZOUFRAS, 2003).

In this paper we focus not on forecasting, but on explaining match outcome. Specifically, we ask the question: what actions during a game characterise winning teams? In contrast to most of the current literature, we neither model actual match result nor goals for and against. We model the shots each team makes during a game. As POLLARD and REEP (1997) noted, a shot on goal by a team can be used as a measure of the effectiveness of a team's possession. By modelling shots using fundamental actions, such as number of passes or crosses, which occur during the course of a match, we provide an insight into which actions improve a teams' effectiveness whilst in possession.

To model shots for and shots against we could employ one of the bivariate Poisson related models used by, for example, GODDARD (2005) when modelling goals. However, one key difference in the relationship between shots for the home team and shots for the away team, in contrast to that between goals for the home team and goals for the away team, is that goals by the two teams display only slight positive or no correlation whereas shots are significantly negatively correlated. It should be noted however, that although goals may not be correlated, there exists some dependence, as demonstrated in DIXON and COLES (1997). The bivariate Poisson models employed to date (GRIFFITHS and MILNE, 1978; BERKHOUT and PLUG, 2004) do not allow for a negative dependence structure, therefore rendering them inappropriate for our purposes. Consequently, it is necessary to seek out an appropriate bivariate discrete distribution, not necessarily having Poisson marginals, which can reproduce not only positive dependence structures, but also negative dependence structures. A natural way forward is to consider copula functions (NELSEN (2006)) to generate various bivariate discrete distributions. It should be noted that extension to the case of multivariate distributions can also be considered.

This paper is structured as follows: Section 2 provides the theoretical underpinnings of copulas and their use in the models employed here. The data we have modelled is then described in Section 3. Section 4 presents the results of fitting the

bivariate distributions to our shots and goals data and Section 5 gives the results of our regression models based on the bivariate distributions. Finally some conclusions are offered in Section 6.

2. Copulas

A copula is a multivariate distribution with all univariate marginal distributions being uniformly distributed on the unit interval, [0,1]; hence C is the distribution of a multivariate uniform random vector. For a bivariate distribution F with margins F_1 and F_2, the copula associated with F is a distribution function $C:[0,1]^2 \to [0,1]$ that satisfies

$$F(x,y) = C\{F_1(x), F_2(y)\}, \quad (x,y) \in \Re^2 . \tag{1}$$

The copula C is uniquely determined on the unit square whenever F_1 and F_2 are continuous. The copula itself characterises the dependence between the random variables X and Y with marginal distributions F_1 and F_2. Thus the copula representation (1) resolves the joint distribution into the marginals F_1 and F_2 and the dependence structure C. When X and Y are discrete random variables taking values on some lattice, Ω, the copula, C, is unique provided $(x,y) \in \Omega$ but not elsewhere; this non-uniqueness is of no consequence however since the region outside Ω is not of interest in the discrete case (NELSEN, 2006). The representation (1) and uniqueness follows essentially from a multivariate extension to the probability integral transformation (JOE, 1997).

Copulas may be constructed in a number of ways. One method that lends itself to the modelling that we consider in this paper is as follows. Let M be a univariate distribution function of a positive valued random variable and let ϕ be its Laplace transform:

$$\phi(s) = \int_0^\infty \exp(-st)\mathrm{d}M(t), \quad s \geq 0 . \tag{2}$$

Now for an arbitrary univariate distribution function F, there exists a unique distribution function G such that

$$F(x) = \int_0^\infty G^w(x)\mathrm{d}M(w) = \phi\{-\log G(x)\} . \tag{3}$$

Rewriting (3) leads to $G(x) = \exp\{-\phi^{-1}(F)\}$. Next, consider the bivariate case and let $G_i(x) = \exp\{-\phi^{-1}(F_i)\}$ for $i = 1,2$. Then

$$\int_0^\infty {G_1}^w(x){G_2}^w(y)\mathrm{d}M(w) = \phi\{-\log G_1(x) - \log G_2 y)\} = \phi\{\phi^{-1}(F_1) + \phi^{-1}(F_2)\}$$

is a bivariate distribution function. The copula is obtained on taking $U(0,1)$ distribution functions for F_1 and F_2:

$$C(u,v) = \phi\{\phi^{-1}(u) + \phi^{-1}(v)\}. \quad (4)$$

This rather simple form is called the Archimedian copula. From (2), a sufficient condition for ϕ to be a Laplace transform family is that it is a convex function such that $\phi(0) = 1$, $\phi(\infty) = 0$. Thus to generate an Archimedian copula, and hence a bivariate distribution, one seeks a generator function ϕ with these properties. Furthermore, a sufficient condition for the resultant copula (4) to have negative dependence is that $\phi^{-1}(e^{-z})$ is concave in *z* (JOE, 1997).

To model a range of dependence, a flexible family of generator functions ϕ_κ, parameterised by a dependence parameter κ, can be used. Two Archimedian copula families that allow negative dependence (for certain values of κ) arise from the generator functions $\phi_\kappa(s) = \kappa^{-1}\log\{1+(e^\kappa - 1)e^{-s}\}$ and $\phi_\kappa(s) = (1-\kappa s)^{1/\kappa}$. These are respectively Frank's (F) copula,

$$C(u,v) = -\kappa^{-1}\log\{1-(1-e^{-\kappa u})(1-e^{-\kappa v})/(1-e^{-\kappa})\}, \quad (\kappa \in \Re), \quad (5)$$

and Kimeldorf and Sampson's (KS) copula

$$C(u,v) = \max\{(u^{-\kappa} + v^{-\kappa} - 1)^{-1/\kappa}, 0\}, \quad (\kappa \in \Re). \quad (6)$$

NELSEN (2006) refers to this latter copula as the Clayton copula.

For any copula, $C(u,v) \le C(u,1) = u$ and $C(u,v) \le C(1,v) = v$ (all $0 \le u,v \le 1$), and so $C(u,v) \le \min(u,v) = M(u,v)$. The copula $M(u,v)$ is called the Frechet upper bound and can be interpreted as the copula with maximum positive dependence (Figure 1a). Furthermore, $C(u,v) \ge \max(u+v-1,0) = W(u,v)$ (all $0 \le u,v \le 1$); $W(u,v)$ is the Frechet lower bound, the copula with maximum negative dependence (figure 1c). The copula families (5) and (6) are comprehensive: that is, they include the Frechet upper and lower bounds. Further, the so-called independence copula, $\Pi(u,v) = uv$ (Figure 1b), with generator function $\psi_{ind} = s\log s$, is obtained as $\kappa \to 0$.

INSERT FIGURE 1 a, b and c HERE.

It is useful to measure the coverage of a copula family in terms of a standard measure of dependence, such as Kendall's τ, since not all copulas are comprehensive. Note that because Kendall's τ is based on ranks and is therefore invariant to a strictly increasing transformation of the margins, its properties depend only on the copula of the bivariate distribution. Furthermore $\tau \in [-1,1]$ and $\tau = -1$ for *W*, $\tau = +1$ for *M*, and $\tau = 0$ for Π (the Frechet lower and upper bounds, and independence copula). For an Archimedian copula, Kendall's τ takes the simple form

$$\tau = 4\int_0^1 \psi_\kappa(s)ds + 1,$$

where $\psi_\kappa(s) = \phi_\kappa^{-1}(s)/\{d\phi_\kappa^{-1}/ds\}$. For the copula

$$C(u,v) = uv\{1-\kappa(1-u)(1-v)\}^{-1}, \quad (-1 \le \kappa \le 1),$$

for example, with generator $\phi_\kappa(s) = (1-\kappa)/(e^s - \kappa)$, $\tau < 1/3$, and the copula family is not comprehensive. For the Kimeldorf-Sampson copula (6), $\tau = \kappa/(\kappa+2)$, and thus τ takes the full range of values, [-1,1], for this family.

There are a large number of Archimedian copula families (NELSEN, 2006, p.116). The choice of a copula family can be guided by the (dependence) properties of that family. Some properties are as follows. Copulas may be reflection symmetry—if specification of the joint survival function in terms of the copula gives rise to the same distribution as specification in terms of the distribution function, then the copula is reflection symmetric. Copulas may be: comprehensive or otherwise; extendable to more than 2 dimensions. They may also exhibit: varying degrees of upper and lower tail dependence; only negative or positive dependence structure. Copula families may be specified by more than one parameter in order to model dependence structure in more detail. NELSEN (2006) and JOE (1997) discuss a number of Archimedian copulas generated by multi-parameter Laplace transform families.

We concentrate here on Frank's copula (6) and Kimeldorf and Sampson's copula (7) for specifying bivariate discrete distributions. For example, a bivariate Poisson distribution with Frank's copula is given by

$$F_{XY}(x,y) = -\frac{1}{\kappa}\log\left(1 - \frac{\left\{1-\exp\left(-\kappa\sum_{i=1}^{x}\frac{e^{-\mu_1}\mu_1^i}{i!}\right)\right\}\left\{1-\exp\left(-\kappa\sum_{j=1}^{y}\frac{e^{-\mu_2}\mu_2^j}{j!}\right)\right\}}{\left(1-e^{-\kappa}\right)}\right) \qquad (7)$$

where $x, y = 0,1,\ldots$, $\mu_1, \mu_2 > 0$, and $-\infty < \kappa < \infty$. This distribution has marginal means μ_1 and μ_2, a dependence parameter κ and negative dependence for $\kappa < 0$. The marginal distributions are independent when $\kappa = 0$. Bivariate geometric and negative binomial distributions may be obtained in a similar manner by replacing u and v in (5) and (6) by the appropriate marginal distribution functions. Using negative binomial distributions as the marginals results in a flexible bivariate distribution which can capture over or under dispersion in the margins, and dependence. Together with the bivariate Poisson distribution (7) and it's equivalent but with the Kimeldorf-Sampson copula, these distributions form the focus of the rest of this paper. We have also fitted bivariate geometric distributions to the data, but results suggest a poor fit in comparison with the alternatives. The parameterisation for the negative binomial distribution used here is given by

$$f\left(x;\mu,\sigma\right) = \frac{\mu^x}{x!}.\frac{\Gamma\left(\sigma+x\right)}{\Gamma\left(\sigma\right)\left(\mu+\sigma\right)^x}.\left(1+\frac{\mu}{\sigma}\right)^{-\sigma}$$

where μ is the mean and σ is a scale parameter.

We thus consider four bivariate distributions generated from two copulas and two pairs of marginal distributions, denoted by F_{PP}, F_{nbnb}, KS_{PP} and KS_{nbnb}, representing

Frank's copula with Poisson marginals, Frank's copula with negative binomial marginals, Kimeldorf and Sampson's copula with Poisson marginals and Kimeldorf and Sampson's copula with negative binomial marginals, respectively.

Given a bivariate discrete distribution specified in terms of marginal distributions $F_1(x;\boldsymbol{\theta}_1)$ and $F_2(y;\boldsymbol{\theta}_2)$ and copula $C(u,v;\kappa)$, the likelihood function for the parameters $(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\kappa)$ given a datum (x_i,y_i) is

$$L\{(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\kappa),(x_i,y_i)\} = \Pr(X=x_i,Y=y_i) = C\{F_1(x_i),F_2(y_i)\} - C\{F_1(x_i-1),F_2(y_i)\} - C\{F_1(x_i),F_2(y_i-1)\} + C\{F_1(x_i-1),F_2(y_i-1)\}.$$

For sample data (x_i,y_i), $i=1,...,n$, the log-likelihood, $\sum_i \log L\{(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\kappa),(x_i,y_i)\}$, may be maximised in a standard way. This approach therefore opens to model fitting a wide and rich field of discrete bivariate distributions with general dependence structure.

3. Data

We have obtained data on 1048 soccer matches in the English Premier League during the period spanning August 2003 to March 2006. The data were collected by the Press Association for use in calculating the official player ratings system of the English Premier League and Championship, the Actim Index. The majority of the data is collected live by observers at the match, and the remainder is collected using post-match video analysis. For each game we have the following variables for the home and away teams (subscripted H and A respectively): goals (g_H, g_A), shots (s_H, s_A), tackles won (t_H, t_A), blocks (b_H, b_A), clearances (cl_H, cl_A), crosses (cr_H, cr_A), dribbles (d_H, d_A), passes made (p_H, p_A), interceptions (i_H, i_A), fouls (f_H, f_A), yellow cards (y_H, y_A) and red cards (r_H, r_A).

4. Results of fitting bivariate distributions to shots

As discussed in the introduction, goals for and against have no significant correlation whilst shots for and against have significant negative correlation, see Table 1.

Table 1: Correlations for shots and goals

Correlation	Corr(s_H, s_A)	Corr(g_H, g_A)
Pearson	-0.269 (0.000)	-0.003 (0.931)
Kendall's τ	-0.191 (0.000)	0.002 (0.924)
Spearman ρ	-0.265 (0.000)	0.003 (0.928)

p-values in parenthesis

Previous studies have used bivariate Poisson models for goals, allowing for positive dependence. However, our data suggest no evidence of positive correlation. The key finding is the negative correlation between home team and away team shots. Table 2 summarises the fits of each of the four bivariate models for shots, with parameter estimates, the corresponding standard errors, the log-likelihood (LL) and the Akaike Information Criterion (AIC) all being shown.

Table 2: Summary results for bivariate models fitted to shots data

Parameter	F_{PP}	KS_{PP}	F_{nbnb}	KS_{nbnb}
κ	-1.367 (0.009)	-0.073 (0.012)	-1.704 (0.196)	-0.123 (0.016)
μ_1 (s_H)	12.696 (0.009)	12.686 (0.009)	12.743 (1.011)	12.750 (1.011)
μ_2 (s_A)	9.603 (0.010)	9.595 (0.010)	9.638 (1.014)	9.645 (1.012)
σ_1			17.802 (1.112)	18.097 (1.113)
σ_2			11.293 (1.101)	11.364 (1.102)
LL	-6205.074	-6231.897	-5996.162	-6019.770
AIC	12416.148	12469.794	12002.324	12069.540

Standard errors are shown in parenthesis.

As can be seen from Table 2, Frank's copula with negative binomial marginals provides the best fit to the shots data. The parameters μ_1 and μ_2 give the fitted marginal means for home team shots and away team shots respectively. The actual means are 12.734 and 9.623, suggesting very reasonable fits.

We now consider diagnostics plots. For an Archimedian copula, $C_\kappa(u,v) = \phi_\kappa\{\phi_\kappa^{-1}(u) + \phi_\kappa^{-1}(v)\}$, GENEST and RIVEST (1993) show that

$$K(s) = \Pr\{C_\kappa(u,v) \le s\} = s - \phi_\kappa^{-1}(s)/\{d\phi_\kappa^{-1}/ds\}. \qquad (8)$$

$K(s)$ is called the *Kendall distribution function* of the copula C_κ. Equation (8) implies that C_κ is uniquely determined by the 1-dimensional projection $\psi_\kappa(s) = \phi_\kappa^{-1}(s)/\{d\phi_\kappa^{-1}/ds\}$. Thus, to consider the fit of a copula, one can compare $\psi_\kappa(s)$ with its empirical estimator (sample equivalent). The sample equivalent of the copula *C* is the proportion of observations in the sample that are less than or equal to (x_i, y_i), component wise, for all $i = 1,...,n$, denoted by $C_n(x_i, y_i)$ $(i = 1,...,n)$. Since dependence between variables *X* and *Y* is completely characterised by the copula *C*, this 1-dimensional projection is independent of the marginal distributions and we may proceed as if the marginal distributions were U(0,1). Letting

$$s_i = \#\{(x_j, y_j) : x_j \le x_i, y_j \le y_i)\}/(n-1), \qquad (9)$$

it follows that $\#(i : s_i \le s)/n, \quad (0 < s < 1)$, is an empirical estimator of $K(s)$, which in turn implies using $\psi_n(s) = s - \#(i : s_i \le v)/n, \quad (0 < s < 1)$, as an empirical estimator of $\psi_\kappa(s)$. The sampling properties of $\psi_n(s)$ are investigated by GENEST and RIVEST

(1993). They show that $\psi_n(s)$ is a consistent estimator of $\psi_\kappa(s)$, and determine an explicit expression for the sampling variance of $\psi_n(s)$ when *C* is from the Kimeldorf-Sampson family of copulas. This variance is given by

$$\operatorname{var}\{\psi_n(s)\} = \frac{\left[\left(s-\psi(s)\right)\left(1-s+\psi(s)\right)+\left(1-\psi'(s)\right)\left(\left(1-\psi'(s)\right)R(s)-2s\left(1-s+\psi(s)\right)\right)\right]}{n} \tag{10}$$

where

$$R(s) = \frac{2\kappa s\left\{\kappa\left(2-s^\kappa\right)^{2-1/\kappa}+\left(1-s^\kappa\right)\left(1-2\kappa\right)-\kappa\right\}}{\left(1-\kappa\right)\left(1-2\kappa\right)\left(1-s^\kappa\right)^2}$$

A simple diagnostic procedure then applies: plot $\psi_n(s)$ $(0<s<1)$, along with approximate confidence bands

$$\psi_n(s) \pm z\sqrt{\operatorname{var}\psi_n(s)} \tag{11}$$

for the unknown $\psi_\kappa(s)$ with $\operatorname{var}\psi_n(s)$ given by (10) (for suitably chosen $z=1.96$, say). The closed form expression for $\operatorname{var}\psi_n(s)$ for the Kimeldorf-Sampson family provides a convenient approximation for other copula families. The maximum likelihood estimator of $\psi\left(s;\hat{\kappa}\right)$ ($\psi_\kappa(s)$ with κ set to its maximum likelihood estimate) may then plotted and compared with the envelope (11) for the unknown $\psi_\kappa(s)$.

For Frank's copula, we have $\phi_\kappa(s) = \kappa^{-1}\log\{1+(e^\kappa-1)e^{-s}\}$, whence

$$\phi^{-1}\left(s;\hat{\kappa}\right) = -\log\left(\frac{1-e^{-s\hat{\kappa}}}{1-e^{-\hat{\kappa}}}\right)$$

and

$$\frac{d\phi^{-1}}{ds} = \frac{-\lambda e^{-s\hat{\kappa}}}{1-e^{-s\hat{\kappa}}}.$$

For discrete data, equation (9) presents a difficulty however due to the likely large number of ties, and resultant distortion of the plots. We propose a simple solution: add a uniform random number, $\mathrm{U}(-0.5, 0.5)$, independently to each margin of the observed data to produce a corresponding pseudo-continuous dataset, and proceed as for the continuous case. Figure 2 shows the diagnostic plot for both the F_{nbnb} and KS_{nbnb} fitted models.

INSERT FIGURE 2 HERE

Also plotted is the empirical estimate, ψ_n with upper and lower 95% confidence intervals (equation 10) and the curve for the independent copula, $\psi_{ind} = s\log s$. As can be seen from the figure, the fitted Kimeldorf-Sampson copula lies outside the

confidence interval whilst the fitted Frank copula lies comfortably inside. The independent copula lies even further outside the confidence intervals.

An alternative procedure is rather than calculate a confidence interval on $\psi_n(s)$, calculate a confidence interval for $\psi(s;\hat{\kappa})$ using the delta method. Thus we have,

$$\text{var}\left\{\psi\left(s;\hat{\kappa}\right)\right\} = \left(\frac{d\psi}{d\hat{\kappa}}\right)^2 \text{var}\left\{\hat{\kappa}\right\},$$

where

$$\frac{d\psi}{d\hat{\kappa}} = \frac{se^{s\hat{\lambda}}}{\hat{\kappa}} + \frac{1}{\hat{\kappa}^2}\left(1-e^{s\hat{\kappa}}\right)\log\left(\frac{1-e^{-s\hat{\kappa}}}{1-e^{-\hat{\kappa}}}\right) + \frac{s}{\hat{\kappa}} + \frac{\left(1-e^{s\hat{\kappa}}\right)}{\hat{\kappa}\left(1-e^{\hat{\kappa}}\right)},$$

with $\text{var}\left\{\hat{\kappa}\right\}$ calculated from the Hessian matrix from the estimation procedure. Figure 3 shows $\psi(s;\hat{\kappa})$ for the F_{nbnb} model, an approximate 95% confidence interval for $\psi(s;\hat{\kappa})$ as calculated using the delta method, and the empirical estimate, $\psi_n(s)$. Again, it is clear that the F_{nbnb} model and the data are in agreement, suggesting the model is more than appropriate for the data.

INSERT FIGURE 3 HERE.

Finally, Figure 4 shows a scatter plot for the pseudo-continuous shots data and contour and surface plots of the fitted F_{nbnb} distribution. The two plots seem very much in agreement. Note that the distribution is not continuous as the surface plot would suggest; the plot is included for illustrative purposes only

INSERT FIGURE 4 HERE.

5. Regression models

A natural way forward is to use covariates in a link function for the mean term of the marginal distributions to produce regression models. Such regression models have been estimated for each copula/marginal distribution pair, and again we find that the F_{nbnb} model provides the best fit to our data. We regress the two marginal means on covariates, as

$$\log\left(\mu_{ij}\right) = x_{ij}\beta_i \quad i = 1,2,$$

where *i* denotes the home and away, *j* represents the j^{th} observation, *x* is a row vector of covariates for the j^{th} observation and β_i is a column vector of regression coefficients to be estimated.

The model parameters have been estimated by maximising the log-likelihood, using various routines written in R (R DEVELOPMENT CORE TEAM, 2005). A pseudo-backward stepwise regression method was adopted to provide the best fitting model. For example, if we consider the home team, only variables over which they could have reasonable control were included in the original model, i.e. if the team could make a conscious effort to do more of the action, then the variable was included as a covariate for that team's mean shots. Thus the following variables were not included in the model, because a team only does these actions in response to the opposition's actions: blocks, clearances and interceptions. The remaining variables were all used as covariates in the first estimation. The least significant variable was then removed and the model refitted. Among the variables dropped were dribbles, yellow and red cards, probably because these events occur relatively infrequently in the data. Further, with regards to red cards, although experience tells us they have an impact on a game, they are rarely given early enough in a game to allow the team with the extra man to take full advantage and create many more shooting opportunities. The final model is given in Table 3.

In comparison to the independent model, $\kappa = 0$, a likelihood ratio test gives a test statistic of 4.208, significant at the 95% level. In addition the AIC also suggests the dependence model is better than the independent model. The parameter estimates have the expected signs, although it is interesting to note that tackles won have a negative impact on that team's number of shots. This is not altogether unexpected, as a team which is forced to make many tackles is less likely to have possession of the ball, suggesting the opposition are the dominant force in the game. One can also examine the relative contributions to shots. For example, if one were assessing two players' performances, a cross by one player could be considered to be equivalent to around 10 passes, in terms of the contribution of these actions to their teams' shots on goals. The key findings are:

- Away team crosses are more likely to be converted to a shot. This may be a consequence of home teams typically adopting a more attacking strategy than away teams.
- Similarly, passes by the away team contribute more to shots than passes by the home team.
- Fouls called against the home team have a greater negative impact on shots than for the away team.

Table 3: Estimated parameters for covariate model

Model	F_{nbnb} covariate model	Independent covariate model
Parameter	Estimate	Estimate
σ_1	3.485 (0.177)	3.508 (0.183)
β_{0H}	2.291 (0.070)	2.249 (0.069)
cr_H	0.011 (0.001)	0.011 (0.001)
p_H	0.001 (0.000)	0.001 (0.000)
t_H	-0.002 (0.001)	-0.002 (0.001)
f_H	-0.011 (0.003)	-0.010 (0.003)
σ_2	3.078 (0.154)	3.113 (0.164)
β_{0A}	1.848 (0.076)	1.766 (0.076)
cr_A	0.017 (0.002)	0.017 (0.002)
p_A	0.002 (0.000)	0.002 (0.000)
t_A	-0.005 (0.002)	-0.005 (0.002)
f_A	-0.005 (0.003)	-0.003 (0.003)
κ	-0.351 (0.207)	
Log-likelihood	-5812.043	-5814.147
AIC	11650.086	11652.294

Our results would seem to provide evidence that the beautiful game works – more passes and crosses contribute to more successful periods of ball possession, i.e. result in a shot. And in answer to the original question posed, what actions characterise successful teams? More passes and crosses do. This result may also suggest the infamous “long-ball” approach to football employed by some teams is less effective in producing shooting opportunities, since, by its very nature, the long-ball has reduced the number of passes.

6. Closing remarks

Copulas provide a rich source of models for multivariate data. Two particular copula families have been used in this paper to generate bivariate Poisson and negative binomial distributions with flexible dependence structure. These discrete distributions have proved here to be very effective in capturing the negative dependence found in shots data for soccer matches. Frank’s copula with negative binomial marginals has been shown to be the most appropriate model for the shots data, with and without covariates that relate to within-match events. The covariate model suggests that playing the “beautiful game” is an effective strategy—more passes and crosses contribute to more effective play (more shots)—and that successful teams are characterised by more

passing and crossing. With within-match events recorded by player, the covariate model could also be used to assess the relative contributions of players to match outcome measures that relate to successful play, and thus form the basis of a system for the ranking outfield players. It might also be developed in order to predict match results, taking account of team selections and players positional roles.

Copula functions may have many other interesting applications in future studies of sports' issues. For instance, copulas could be employed for modelling results in a tournament or a series of annual contests, say. In the latter case, bivariate Bernoulli distributions with general dependence structure might be used to model serial dependence in a sequence of matches of the kind considered by BAKER and SCARF (2006). Similarly, multivariate Bernouilli distributions might be used to model dependence structure in the outcomes of a round-robin such as the Six-Nations rugby tournament or the group-stages of a sports tournament—such a model might be used to predict the outcomes of matches in the latter part of a group-stage given the results in the group-stage to date.

Acknowledgements

The authors would like to thank their colleagues Rose Baker, David Forrest, Dan Jackson, and Martin Newby for useful discussions about the data and models. We are also grateful PA Sport for making the soccer data available to us.

References

BAKER, R.D., P.A. SCARF (2006), Predicting the outcomes of annual sporting contests. *Applied Statistics*, 55, 225-239.

BERKHOUT, P., E. PLUG (2004), A bivariate Poisson count data model using conditional probabilities. *Statistica Neerlandica*, 58, 349-364.

DIXON, M.J., S.G. COLES (1997), Modelling association football scores and inefficiencies in the football betting market. *Applied Statistics*, 46, 265-280.

DOBSON S., J. GODDARD (2001), *The Economics of Football*. Cambridge University Press.

GENEST, C., L-P. RIVEST (1993), Statistical inference procedures for bivariate Archemdian copulas, 88, 1034-1043. *Journal of the American Statistical Association.*

GODDARD, J. (2005), Regression models for forecasting goals and match results in association football. *International Journal of Forecasting*, 21, 331-340.

GRIFFITHS, R.C., R.K. MILNE (1978), A class of bivariate Poisson processes. *Journal of Multivariate Analysis*, 8, 380-395.

JOE H. (1997), Multivariate Models and Dependence Concepts. Chapman and Hall, London.

KARLIS D., I. NTZOUFRAS (2003), Analysis of sports data using bivariate Poisson models. *The Statistician*, 52,381-393.

KONING, R.H. (2000), Balance in competition in Dutch soccer. *The Statistician*, 49, 419-431.

MAHER, M.J. (1982), Modelling association football scores. *Statistica Neerlandica*, 36, 109-118.

NELSEN, R.B. (2006), *An Introduction to Copulas, 2nd Edition*. Springer, New York.

POLLARD R., C. REEP (1997), Measuring the effectiveness of playing strategies at soccer. *The Statistician*, 46, 541-550.

R DEVELOPMENT CORE TEAM (2005), R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

SZYMANSKI S. (2003), The economic design of sporting contests. *Journal of Economic Literature*, 41, 1137-1187.

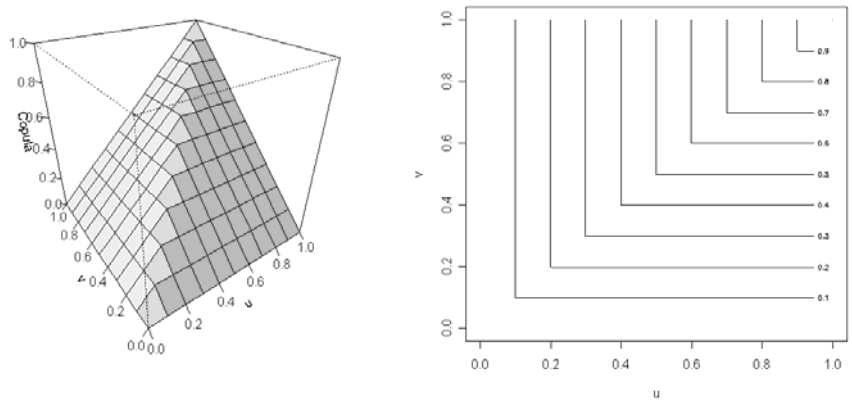

Figure 1a: Surface and contour plot of the Frechet upper bound, $M(u,v)$.

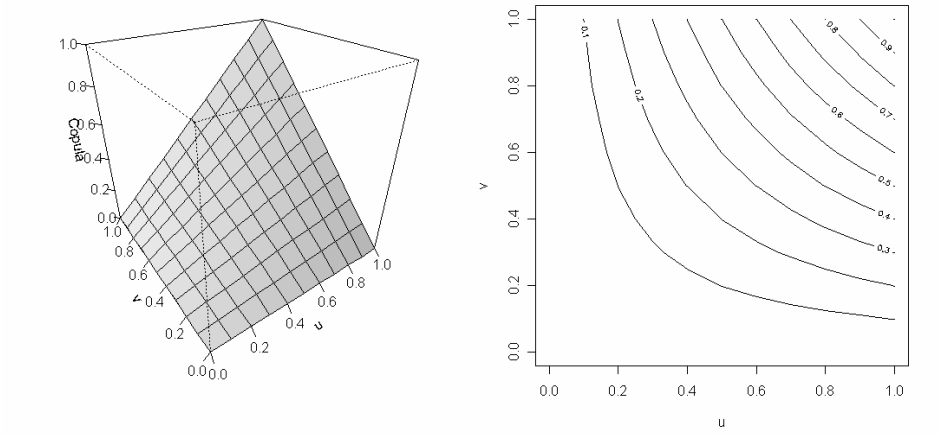

Figure 1b: Surface and contour plots for the independence copula, $\prod(u,v)$.

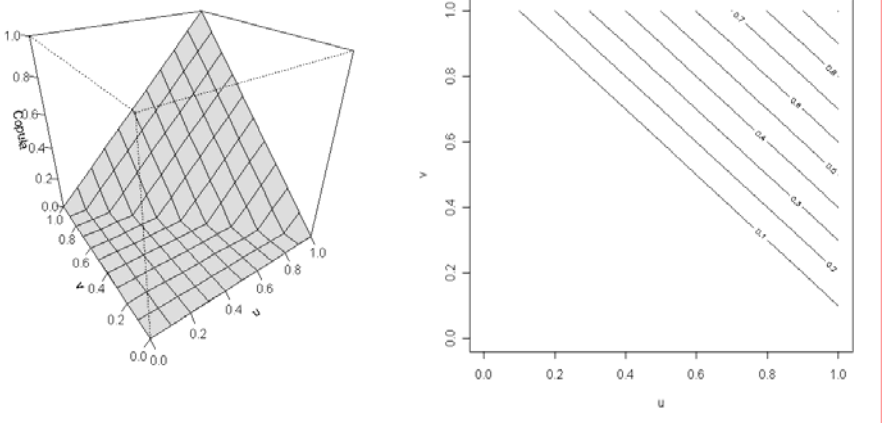

Figure 1c: Surface and contour plot of the Frechet lower bound, $W(u,v)$.

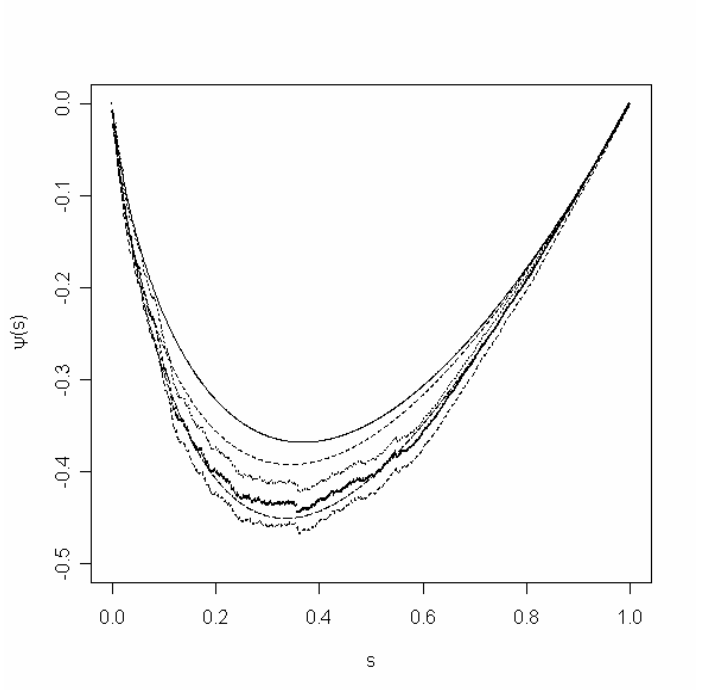

Figure 2: Diagnostic plot for copula fit: solid line = ψ_{ind}, bold dotted=$\psi_n(s)$, dotted=approx. 95% confidence interval for $\psi_n(s)$, dashed = $KS_{nbnb}\ \psi(s;\hat{\kappa})$, long dash = $F_{nbnb}\ \psi(s;\hat{\kappa})$

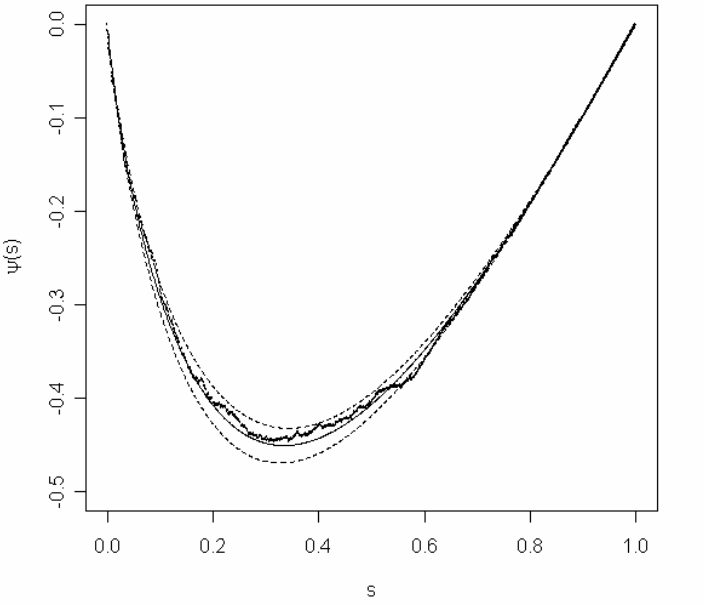

Figure 3: Diagnostic plot for copula fit: solid line = $F_{nbnb}\ \psi(s;\hat{\kappa})$, bold dotted = $\psi_n(s)$, dashed = approx. 95% confidence interval from delta method for $\psi(s;\hat{\kappa})$

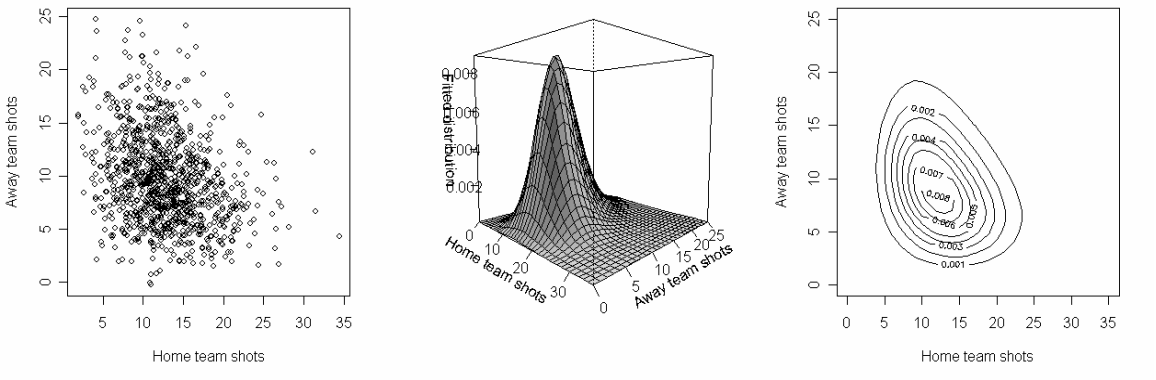

Figure 4: Scatter plot of pseudo-continuous shots data and fitted F_{nbnb} distribution, surface and contour plots.