A Bayesian network to analyse basketball players' performances: a multivariate copula-based approach

Pierpalo D'Urso[1] · Livia De Giovanni[2] · Vincenzina Vitale[1]

Abstract

Statistics in sports plays a key role in predicting winning strategies and providing objective performance indicators. Despite the growing interest in recent years in using statistical methodologies in this field, less emphasis has been given to the multivariate approach. This work aims at using the Bayesian networks to model the joint distribution of a set of indicators of players' performances in basketball in order to discover the set of their probabilistic relationships as well as the main determinants affecting the player's winning percentage. From a methodological point of view, the interest is to define a suitable model for non-Gaussian data, relaxing the strong assumption on normal distribution in favour of Gaussian copula. Through the estimated Bayesian network, we discovered many interesting dependence relationships, providing a scientific validation of some known results mainly based on experience. At last, some scenarios of interest have been simulated to understand the main determinants that contribute to rising in the number of won games by a player.

Keywords Bayesian networks · Gaussian copula · Basketball data · Four factors · Multivariate dependence

1 Introduction

Sports analytics has experienced rapid growth both for teams and individual players in recent years. Statistics tools can be particularly useful to assess performances and define efficient gaming strategies. In this paper, we focus on basketball, a sports pioneer in using analytic tools, and data from the National Basketball Association (NBA).

✉ Vincenzina Vitale
vincenzina.vitale@uniroma1.it

Pierpalo D'Urso
pierpaolo.durso@uniroma1.it

Livia De Giovanni
ldegiovanni@luiss.it

1   Department of Social Sciences and Economics, Sapienza University of Rome, P.za Aldo Moro 5, 00185 Rome, Italy

2   Department of Political Sciences, Data Lab Luiss University, Viale Romania, 32, 00197 Rome, Italy

As far as the literature on statistics in basketball is concerned, many of the proposed approaches can be traced back to the Bill James's Sabermetrics approach in baseball (James, 1984, 1987). The seminal work is, however, considered that of Oliver (2004): he identifies the so-called "Four Factors" that affect the game outcome with different weights and they are: the Effective Field Goal percentage, the Turnovers per Possession, the Offensive Rebounding percentage and the Free Throw Attempt rate, analysed later with more detail.

Kubatko et al. (2007) give a common starting point for research in basketball by providing a detailed overview and description of the main sources of data for basketball statistics; they also define the main basic variables used in what described as the "mainstream of basketball statistics", firstly emerged outside the academy and due to Oliver (2004), Hollinger (2004), Hollinger and Hollinger (2005). In Nikolaidis (2015) one can find some indicative ideas for the statistical analysis of basketball data; he also shows that any basketball team can improve significantly its decision-making process if it relies on statistical support.

Yang et al. (2014) evaluate the efficiency of National Basketball Association (NBA) teams under a two-stage DEA (Data Envelopment Analysis) framework. Applying the additive efficiency approach, they decompose the overall team efficiency into first-stage wage efficiency and second-stage on-court efficiency, finding out the individual endogenous weights for each stage.

Other recent scientific proposals to model player and team performances make use of network models (Piette et al., 2011; Fewell et al., 2012; Skinner & Guy, 2015; Xin et al., 2017). In the class of models accounting for spatial features, Metulini et al. (2018) try to identify the pattern of surface area in basketball that improves team performance while Zuccolotto et al. (2019) propose a spatial statistical method based on classification trees to assess the scoring probability of teams and players in different areas of a court map.

Sandholtz et al. (2020) propose a new metric for spatial allocative efficiency: they estimate the player's field goal percentage at every location in the offensive half court using a Bayesian hierarchical model. Then, comparing the estimated field goal percentage with the empirical field goal attempt rate, they identify the areas where the lineup exhibits inefficient shot allocation.

As far as the individual performances are concerned, Fearnhead and Taylor (2011) define a proper statistical methodology to estimating the ability of NBA players by comparing their team performance when they are on the court with respect to when they are off it. Lopez and Matthews (2015) propose an NCAA men's basketball predictive model while is due to Cervone et al. (2016) a multiresolution stochastic process model for predicting basketball possession outcomes. Engelmann (2017) analyses the possession-based player performance too while Wu and Bornn (2018) model the offensive player movement in professional basketball.

Among the regression-based approach, Deshpande and Jensen (2016) propose a Bayesian linear regression model to estimate and evaluate the impact NBA players have on their teams' chances of winning while Sill (2010) shows that the Adjusted Plus-Minus (often abbreviated APM, a basketball analytic used to predict the impact of an individual player on the scoring margin of a game after controlling for the strength of every teammate and every opponent during each minute he's on the court) can be improved by using ridge regression.

Other alternatives based on Bayesian regression models are in Page et al. (2007) and Deshpande and Jensen (2016) while in the work of Casals and Martinez (2013), a mixed model is proposed and applied to model player performances.

In the field of neural networks, one can refer to the works of Loeffelholz et al. (2009), Blaikie et al. (2011), Wang and Zemel (2016), Shen et al. (2020), Babaee Khobdeh et al.

(2021). For an extensive literature on the use of Statistics in basketball, an interesting review is provided by Terner and Franks (2021).

But, as stated by Baghal (2012), there is a lack of attention to applying multivariate analyses on basketball metrics even though it is well known that winning a game depends on a plurality of interrelated variables affecting the game outcome each with a different weight. To this aim, in the same work, Baghal (2012) uses a structural equation model to determine whether the Four Factors can be modelled as indicators of individual latent factors for the offensive and defensive quality. Team salaries are added to the model too in order to estimate the relationship between the cost and offensive and defensive quality.

With the same aim of enlarging the analysis by considering a multivariate approach, we propose the use of Bayesian networks (BNs) (Pearl, 1988), i.e. probabilistic expert systems able to identify complex dependence structures, such as that can exist among the set of performance indicators, and to carry out *what-if* analysis, at the same time.

Based on our knowledge, the use of BN in sports is rather limited to very few works (Razali et al., 2017; Constantinou et al., 2013), the latter using BNs with other scopes.

As deeply discussed later, the attractiveness of a BN mainly stems both from its pictorial representation of the dependency structure through the graph topology and from its inference engine; the latter makes a BN a decision support system since the reasoning (inference) is performed by setting some variables in known states and then updating the probabilities of the remaining variables conditioned by this evidence. The so-called *belief updating* can thus help to identify the most impactful variables and suggest the best strategies in different contexts.

From a methodological point of view, this proposal copes with the problem of handling non-Gaussian distributions, avoiding the bad practice of variables' discretization. Therefore, a suitable extension to the Gaussian copula is considered to estimate the network structure under the above assumption, relaxing Guassianity. Specifically, the Non-Parametric BNs (NPBNs) proposed by Hanea et al. (2006) and Kurowicka and Cooke (2006) are applied in this paper.

Based on all these considerations, this work aims at introducing to the use of BNs in sports analytics; their versatility and their natural vocation to be used for prediction purposes could represent a valid instrument to predict team and player performance, to suggest strategies as well as a valid tool for decision-making in sports.

Specifically, we use BNs to identify the main determinants affecting the player winning percentage taking into account the most important performance indicators suggested in the literature. Moreover, we propose a model that combines information from data and experts which represents an important added value in this context.

Data used in this work have been downloaded from the NBA website and refer to 377 players who played at least 10 games and more than 6 min on the court in the regular season 2019–2020.

The paper is organized as follows. In Sect. 2 we provide a theoretical description of BNs and their structural learning methods followed by a brief discussion about the limits of the discretization; in Sect. 3 we introduce the class of Non-Parametric BNs followed by a little digression on the Nonparanormal estimator proposed by Liu et al. (2009) being the Gaussian copula a specific case of the class of Nonparanormal models. After the description of variables used in this study in Sect. 4, we discuss the estimated model in Sect. 5. The last section includes some comments and conclusions.

2 Basics on Bayesian networks

Bayesian Networks (Cowell et al., 1999; Pearl, 1988) are probabilistic graphical models used to deal with uncertainty in complex probabilistic domains. They bring together graph theory and probability theory to model multivariate dependence relationships among a large collection of random variables. It consists of a qualitative component, in the form of a directed acyclic graph (DAG), and of a quantitative component, in the form of conditional probabilities associated to each node of the DAG. Hence, they can be defined as multivariate statistical models satisfying sets of (conditional) independence statements encoded in a DAG.

More formally, given a random vector $\mathbf{X} = \{X_1, \ldots, X_k\}$ with distribution P, each node in the DAG corresponds to a random variable X_i, resulting in a graph with k vertices and a set of directed edges between pairs of nodes; hence, a DAG G is a pair $G = (V, E)$, where V is the set of vertices (or nodes) and E the set of directed edges. While each node corresponds to a random variable, a missing arrow between nodes imply (conditional) independence.

Specifically, the conditional independence properties of a joint distribution P can be read directly from the graph by applying the rules defined in the so-called d-separation criterion introduced by Pearl (1988) or the directed global Markov criterion due to Lauritzen et al. (1990), that is an equivalent criterion to that of the d-separation.

Based on Pearl's criterion, given three disjoint subsets of nodes $\mathbf{X}, \mathbf{Y}$ and $\mathbf{Z}$ in a DAG G, $\mathbf{Z}$ is said to d-separate $\mathbf{X}$ from $\mathbf{Y}$, if for all paths between $\mathbf{X}$ and $\mathbf{Y}$ there is a vertex w such that either:

– the connection is serial or diverging and $w \in \mathbf{Z}$ (it is equivalent to say that w is instantiated);
– the connection is converging, and neither w nor any of its descendants (i.e. the nodes that can be reached from w) are in $\mathbf{Z}$.

A serial connection is of the type $x \longrightarrow w \longrightarrow y$ (or $x \longleftarrow w \longleftarrow y$), a diverging connection is $x \longleftarrow w \longrightarrow y$ while a converging one is $x \longrightarrow w \longleftarrow y$. Clearly, if $\mathbf{X}$ and $\mathbf{Y}$ are not d-separated, they are d-connected.

In the graph in the Fig. 1, for example, X_3 d-separates X_1 and X_2 from X_4, implying the conditional independence among the same corresponding variables in P, i.e. $(X_1, X_2) \perp X_4 | X_3$.

A directed edge from X_i to X_j states that X_j is influenced by X_i: using graph terminology, X_j is said the *child* of X_i that, in turn, is said the *parent* of X_j. The set of parents of X_j in the graph G is denoted by $pa(X_j)$. In the DAG of Fig. 1, for example, $pa(X_3) = \{X_1, X_2\}$.

Fig. 1 An example of DAG with 5 vertices
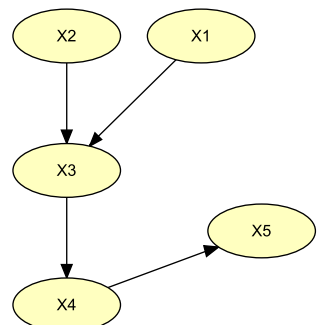
Moreover, with each variable X_i in the DAG, is associated a conditional probability $Pr(X_i|pa(X_i))$; all set of these conditional probabilities constitute the quantitative part of the network.

Based on the local Markov property[1] and applying the chain rule, the joint distribution P over the random vector $\mathbf{X}$ can be factorized as:

$$P(X_1 \ldots X_k) = \prod_{i=1}^{k} P(X_i|pa(X_i)) \tag{1}$$

Besides its ability to decompose a multivariate distribution in a set of local computations, the power of Bayesian networks stems also from the availability of fast and efficient algorithms that perform probabilistic inference (Lauritzen & Spiegelhalter, 1988; Jensen et al., 1990). They update the marginal probability distributions of some variables when the information on one or more distributions of other variables is observed, thus allowing to carry out *what-if* analysis.

The structure of a BN (*i.e* the DAG), as well as its parameters, can be defined by means of the expert knowledge or learned directly from data using *structural learning techniques* (Buntine, 1996; Neapolitan, 2003); in practice, the combination of both approaches is, however, recommended, introducing the prior information as constraints in the first learning phase. The next paragraph provides a little overview of the main methods proposed in the literature to learn the DAG structure.

2.1 Structural learning in BNs

The structural learning techniques for BNs fall into one of these two mainstream approaches: *constraint-based* (Spirtes et al., 1993; Cooper, 1997) and *score-based* (Heckerman, 1995; Cooper & Herskovits, 1992) algorithms.

A DAG learnt by means of a *constraint-based* algorithm entails all the marginal and conditional independence relationships inferred from independence tests: the most famous algorithm in this class is the PC algorithm due to Spirtes et al. (1993).

A DAG learnt by means of a *score-based* approach is the result of an optimization search strategy based on a scoring function. The graph structure is that which corresponds to the optimal score metric (the Penalised Likelihoods, AIC and BIC, are frequently used) among candidate structures, hence, identifying the better fitting network.

It must be noticed that hybrid methods have also been proposed in the literature, combining both *constraint-based* and *score-based* techniques. The two best-known members of this family are the Sparse Candidate algorithm (SC) by Friedman et al. (1999b) and the Max-Min Hill-Climbing algorithm (MMHC) by Tsamardinos et al. (2006).

In this work, we apply a *score-based* algorithm known as the *hill climbing* algorithm, which is a heuristic optimization procedure given that the investigation of the space of all possible DAGs is infeasible in practice by growing exponentially with the number of nodes.

The *hill climbing* uses a greedy search strategy: generally starting from a network structure (more often the empty one), it iteratively performs small changes in the current candidate network by adding, deleting or reversing one arc at a time, stopping when any of changes improves the chosen score function.

In this study, specifically, the *hill climbing* with bootstrap resampling and model averaging is applied as proposed by Friedman et al. (1999a) and implemented in the R package

---

[1] The local Markov property states that each node X_i is conditionally independent of its non-descendants (e.g., nodes X_j for which there is no path from X_i to X_j) given its parents.

bnlearn. More robust to noise in the data, this technique consists of sampling data using bootstrap n times and applying the chosen learning algorithm to each bootstrap sample. Then, the resulting averaged network is a graph whose edges are those with a strength greater than a threshold value (see Scutari and Nagarajan (2011) for details about the computation of the strength significance threshold) where the strength of each arc is defined in terms of the number of times the arc appears over the n learned networks.[2]

Despite the broad fields of applicability, in the parametric setting, most inference and structural learning methods work under the joint multinomial assumption for discrete data and the joint normality for continuous data. In the case of mixed variables, the conditional Gaussian distribution (CLGBNs; Lauritzen and Wermuth (1989)) is assumed with the constraint that continuous nodes can not have discrete children.

Focusing on the continuous case, which is of interest in this work, in the Gaussian BNs (GBNs, Geiger and Heckerman (1994)), a multivariate normal distribution is assumed for $\mathbf{X}$ implying that each component X_i follows a univariate Gaussian distribution too. Therefore,

$$f(X_i|pa(X_i) = \mathbf{x}) \sim N(\alpha_i + \boldsymbol{\beta}^T \mathbf{x}; \sigma_{X_i}^2). \tag{2}$$

Hence, the parameters of the local distributions are the β coefficients of the linear regression model in (2). The influence of the parents on a child is translated in terms of these partial regression coefficients, meaning that their conditioning effect is given by the additive linear term in the mean without affecting the variance.

The main drawback associated with the use of the Gaussian BNs. relates to the joint normality which may be a very strong assumption, very often unrealistic; in the next paragraph, we shall examine in detail this issue.

2.2 The joint normality assumption and its limits

When the marginal distributions of the continuous variables of the domain are far from normal, the common practice consists in their discretization; very often, it results in loss of information and, in some cases, can destroy the true underlying dependence structure. To this aim, Nojavan et al. (2017) designed an experiment to evaluate the impact of the commonly used discretization methods for BNs showing that different methods result in differing future predictions. They state that "we would caution against decisions based on models for which the outputs vary by the choice (of discretization method) that does not have justifiable scientific basis" arguing that, in general, when possible, discretization of continuous variables should be avoided.

As highlighted by Rohmer (2020), a second drawback is computational since one cannot ignore that the size of conditional probability tables grows exponentially as the number of states of the involved nodes increases.

The class of Non-Parametric BNs (NPBNs) proposed in the literature by Kurowicka and Cooke (2006) and Hanea et al. (2006) and based on a D-vine copula approach can overcome these limits. Joint density is modelled using joint normal copula without requiring any parametric assumption on the marginals, as opposed to the parametric case, which justifies the use of the adjective "Non-Parametric".

Other interesting works combining the distributional flexibility of pair-copula constructions (PCCs) with the parsimony of the conditional independence models associated with

---

2 The algorithm also estimates the strength of each arc's directions conditional on the arc being present in the graph.

the DAG, can be found in Bauer et al. (2012), Hobæk Haff et al. (2016), Bauer and Czado (2016), Pircalabelu et al. (2017).

Is due to Harris and Drton (2013) the proposal of the Rank PC Algorithm for Nonparanormal Graphical Models; properly, they extend the well-known PC structural learning algorithm for Gaussian BNs (Spirtes et al., 1993) to the case of Gaussian copula.

Elidan (2010) proposes the Copula Bayesian Networks (CBNs): as in BNs, they make use of a graph to encode independencies but rely on local copula functions and an explicit globally shared parameterization of the univariate densities. Therefore, joint density is constructed via a composition of local copulas and marginal densities. A generalization of the CBNs to hybrid BNs can be found in Karra and Mili (2016).

In the class of the undirected graphical models, Liu et al. (2009) introduce the family of Nonparanormal distributions to relax the Gaussian assumption proposing two different methods of estimation for a nonparametric undirected graphical model.

In our work, under the assumption of a Gaussian Copula, we estimate a NPBN to identify the main determinants affecting the player winning percentage considering several important performance indicators in basketball. We use the theoretical framework of NPBNs proposed by Kurowicka and Cooke (2006) except for the structural learning phase, *i.e* the estimation of the DAG: to learn the graph, we apply the *Hill Climbing* algorithm to a suitable transformation of the original random variables in normal scores as suggested by Liu et al. (2009) and defined more in detail later.

In the next section, we introduce the theoretical framework of NPBNs.

3 Non-parametric Bayesian networks

The Non-Parametric Bayesian Networks, proposed in the literature by Hanea et al. (2006) and Kurowicka and Cooke (2006), are BNs whose dependence structure is modelled through a Gaussian Copula. The term "Non-Parametric" has been referred to the marginals since no distributional assumption is strictly required, i.e. the empirical distributions could be used.

For further interesting readings we refer to Hanea et al. (2010) and Kurowicka and Cooke (2010).[3]

In a general setting, each node corresponds to a quantitative variable with an arbitrary invertible distribution function while the directed edges are associated with (conditional) parent–child rank correlations computed under the assumption of a joint normal copula. These (conditional) copulae, assigned to the arcs of the NPBN, depend on a (non-unique) ordering of the parent nodes. We highlight that the ordering of parent nodes can be assessed by the experts considering the decreasing strength of influence on the child, based on data availability or other qualitative considerations.

The main result arises from the following theorem (Hanea et al., 2006):

- given k variables $X_1, \ldots, X_k$ with invertible distribution functions $F_1, \ldots, F_k$;
- given the corresponding DAG with k nodes entailing all conditional independence relationships;
- given a specification of the conditional rank correlations on the arcs of the NPBN;
- given a copula realizing all correlations $[-1, 1]$ satisfying the so-called zero-independence property[4];

---

[3] For interesting applications see Dalla Valle and Kenett (2015), Vitale et al. (2018), Marella et al. (2019).

[4] If a chosen copula satisfies the so-called zero-independence property, a zero correlation implies the independent copula as well zero partial correlation implies conditional independence.

the joint density over the k variables is (1) uniquely determined, (2) satisfies the BN factorization and (3) the conditional rank correlations are algebraically independent.

Moreover, the *zero independence property* implies that the conditional independence statements can be translated in the graph, assigning zero rank correlation to the corresponding arc, i.e. the arc can be removed.

The NPBNs specification also implies that they require only the definition of the one-dimensional marginal distributions and the arbitrary (conditional) rank correlations associated with the arcs of the BN to be quantified and well specified.

More important, the Gaussian copula has been chosen not only because it satisfies the *zero independence property* but also because it allows to perform conditioning analytically. This allows for extremely fast exact inference, therefore, speeding the probabilistic inference process in complex domains.

More specifically, let X_1 and X_2 be two continuous random variables, with invertible distribution functions F_1 and F_2 while Y_1 and Y_2 are the corresponding transformation to standard normal variables. The conditional distribution of $X_1|X_2$ is obtained directly applying $F_1^{-1}(\Phi(Y_1|Y_2))$.

This inference engine performing conditioning analytically is efficiently implemented in the `UniNet` software.

For sake of completeness, one needs to make a little digression on the validation procedure for NPBNs proposed by Hanea et al. (2006), Hanea et al. (2010) and Joe and Kurowicka (2011) to test the Gaussian Copula assumption.

To validate that the joint normal copula distribution is a valid model for the multivariate data, they defined a new measure of multivariate dependence equal to the determinant of the rank correlation matrix. Lying in [0, 1], it is equal to 1 if all variables are independent, to 0 in the case of multivariate linear dependence. All values in such interval take into account the different degree of dependence.

Two determinants have to be computed: the DER, the determinant of the empirical rank correlation matrix and the DNR, the determinant of the empirical normal rank correlation matrix computed on the transformed variables (that are standard normals) applying the inverse Pearson transformation such that $\rho_s = \frac{6}{\pi} arcsin(\frac{\rho}{2})$ where ρ is the Pearson correlation coefficient while ρ_s the Spearman one.

This validation step is carried out by simulating the sampling distribution of DNR and checking whether DER is within the 90% central confidence band of DNR; if yes, the Gaussian Copula assumption cannot be rejected at the 10% significance level. This validation procedure is also used in this work.

The DAG selection in `Uninet` can be done by removing from the saturated graph those edges with a small conditional rank correlation[5]; if the DNR falls within the 90% central confidence band of the determinant of the rank correlation matrix (computed under the assumption of joint normal copula) of the selected non-saturated BN, the model adequately approximates the complete graph and could be retained.

Since in the above model selection the arcs' directions are assigned based on non-statistical considerations, such as subject-matter knowledge and time/logical ordering, in this paper, the structural learning phase is replaced by the following procedure: the *hill climbing* algorithm with bootstrap resampling and model averaging is applied to a suitable transformation in normal scores of the original variables as proposed by Liu et al. (2009) and defined in the next paragraph.

---

[5] Another model selection procedure starts from a BN with no arcs, adding all the edges whose rank correlation (in the normal rank correlation matrix) is greater than the given threshold.

We point out that the parameters' learning phase is based on the procedure described for the NPBNs, thus based on the conditional rank correlations.

3.1 The nonparanormal estimator proposed by Liu et al. (2009)

Following Liu et al. (2009), the Nonparanormal models extend Gaussian models to semi-parametric Gaussian copula models, with nonparametric marginals. Formally:

Definition 1 a random vector $\mathbf{X} = (X_1, \ldots, X_k)^T$ has a Nonparanormal distribution if there exist functions $\{f_j\}_{j=1}^k$ such that $\mathbf{Z} \equiv f(\mathbf{X}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $f(\mathbf{X}) = (f_1(X_1), \ldots, f_k(X_k))$. Then, $\mathbf{X} \sim NPN \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}, f)$.

In another equivalent way, let $\{f_j\}_{j=1}^k$ be a set of monotone univariate functions and let $\boldsymbol{\Sigma}$ be a positive-definite correlation matrix. Then a k-dimensional random variable $\mathbf{X} = (X_1, \ldots, X_k)^T$ has a Nonparanormal distribution $\mathbf{X} \sim NPN \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}, f)$ if $f(\mathbf{X}) := (f_1(X_1), \ldots, f_k(X_k)) \sim N_k(0, \boldsymbol{\Sigma})$.

Therefore, a random vector $\mathbf{X}$ belongs to the Nonparanormal family if there exists a set of monotonically increasing transformations functions $\{f_j\}_{j=1}^k$ such that $f(\mathbf{X}) = (f_1(X_1), \ldots, f_k(X_k))^T$ is Gaussian.

Moreover, as a lemma, Liu et al. (2009) proved that the Nonparanormal distribution $NPN(\boldsymbol{\mu}, \boldsymbol{\Sigma}, f)$ is a Gaussian copula when the f_j's are monotone and differentiable.

A little digression is needed to remember that, in the classical Gaussian Graphical Models, one assumes that the observations have a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$, and covariance matrix $\boldsymbol{\Sigma}$. More importantly, the conditional independence can be implied by the inverse covariance (concentration) matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$. Indeed, if $\Omega_{jk} = 0$, then the i-th variable and j-th variable are conditional independent given all other variables. In other words, the graph G is encoded by the precision matrix.

The following Lemma (Liu et al., 2009) actually states that it also holds for the Nonparanormal graphical models.

Lemma 1 *If* $\mathbf{X} \sim NPN(\boldsymbol{\mu}, \boldsymbol{\Sigma}, f)$ *is Nonparanormal and each* f_j *is differentiable, then* $X_i \perp\!\!\!\perp X_j | X_{\setminus\{i,j\}} \Leftrightarrow \Omega_{ij} = 0$, *where* $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$.

The important point to note is that, although the NPN identifies a more flexible and broader class of distributions than the parametric one, the independence relations among the variables are still encoded in the precision matrix $\boldsymbol{\Sigma}^{-1}$.

For the scopes of Liu et al. (2009), the primary objective of the Nonparanormal model (NPN) is to estimate the underlying sample covariance matrix for a better recovery of the underlying undirected graph. Therefore, the random variable $\mathbf{X} = (X_1, \ldots, X_k)$ is replaced by the transformed random variable $f(\mathbf{X}) = (f_1(X_1), \ldots, f_k(X_k))$ assuming that $f(\mathbf{X})$ is multivariate Gaussian.

Based on the above theoretical framework, the Nonparanormal estimator proposed by Liu et al. (2009) is based on a two-step procedure: (1) replace the observations, for each variable, by their respective normal scores, i.e. they estimate the transformation functions; (2) apply the graphical lasso to the transformed data to estimate the undirected graph.

In the step (1), as a Non-parametric estimator for the functions $\{f_j\}$, Liu et al. (2009) proposed the normal score-based Nonparanormal estimator, i.e. a truncated empirical distribution (the so called Winsorized estimator) whose computation is done by using the R package `huge`.

In detail, let $x_1, \ldots, x_n$ be n observations of the X variable and let $u_1, \ldots, u_n$ be the corresponding rank; denoting with Φ the normal CDF function, the truncated normal score f_i is estimated by:

$$\hat{f_i} = \begin{cases} \Phi^{-1}(\delta) & if \ \hat{u}_i \leq \delta \\ \Phi^{-1}(\hat{u}_i) & if \ \delta < \hat{u}_i \leq 1 - \delta \\ \Phi^{-1}(1 - \delta) & if \ \hat{u}_i > 1 - \delta \end{cases}$$

where $\hat{u}_i = \frac{u_i}{n+1}$ and $\delta = \frac{1}{4n^{1/4}\sqrt{\pi log(n)}}$.

For further mathematical details, see Liu et al. (2009, 2012).

Following Liu et al. (2009), we replace the observations with their respective normal scores; in this way, we can apply to them the *Hill climbing* algorithm as for the Normal case.

In the next sections, we describe the variables used in the analysis and, then, apply the proposed learning strategy to them.

4 Data description

The variables used in this study were downloaded from the NBA website and listed in Table 1. They refer to 377 players, all those who played at least 10 games and at least 6 min per game in the 2019–2020 regular season.

Thanks to Oliver's studies (Oliver, 2004), in the Basketball, particular attention has been given to the so-called "Four Factors": the effective Field Goal percentage, the Turnovers per Possession, the Offensive Rebounding percentage and the Free Throw Attempt rate.

These measures are considered good proxies for the overall offensive or defensive performance and are, in general, not all equivalent. Indeed, in his studies, Oliver assigned a weight to each factor so that the effective Field Goal percentage is the most important followed by the turnovers, the rebounding, and the free throws.

Because of their importance, they have included in our work. In detail, the effective field goal percentage (EFG) is the field goal percentage adjusting for made 3-point field goals, i.e.

Table 1 Basketball variables

Variable	Label
Age	Age
Assist percentage	AST
Assist ratio	AST_RATIO
Effective field goal percentage	EFG
Free throw rate	FTA_RATE
Defensive rating	DEFRTG
Defensive rebound percentage	DREB
Minutes played	MIN
Turnover ratio	TO_RATIO
Offensive rebound percentage	OREB
Offensive rating	OFFRTG
Number of possession per 48 min	PACE
Usage percentage	USG
Wins over played games	WIN

accounts for made three-pointers and serves to catch a player's (or team's) shooting efficiency from the field.

The turnovers per possession (To_Ratio) is the number of turnovers a player (or a team) averages per 100 possessions used.

The Offensive Rebounding percentage (OREB) is the percentage of available offensive rebounds a player obtains while on the floor; players, as well as teams, by obtaining an offensive rebound, can have one more attempt to make a basket that, in turn, can rise the winning probability.

The Free Throw Attempt rate (FTA_RATE) is the number of free throws made by a player in comparison to the number of field goal attempts; it expresses at the same time the player's ability to get to the foul line and to make foul shots.

In addition, the Defensive Rebound percentage (DREB) is also considered, that is the percentage of available defensive rebounds a player obtains while on the floor.

We take into account the offensive and defensive ratings (OFFRTG and DEFRTG) too; they could be seen as measures of the player's efficiency per possession. In particular, OFFRTG is the number of points per 100 possessions that the team scores while that individual player is on the court while DEFRTG is the number of points per 100 possessions that the team allows while that individual player is on the court.

We also include the Assist percentage (AST), the percentage of a team's assists that a player has while on the court and the Assist Ratio (AST_RATIO), i.e. the number of assists a player averages per 100 possessions used.

Moreover, we take into account the age of the player (AGE), the Minutes played by a player (MIN) as well the number of possession per 48 min (PACE) and the usage percentage (USG), i.e. the percentage of team plays used by a player when they are on the floor. At last, the variable that refers to the wins over the played games (WIN) is also included.

5 The estimated network

Based on the previously described variables and 1.000 simulations, we cannot reject, at the 10% level, the hypothesis that data were generated from the joint normal copula since the determinant of the empirical rank correlation matrix (DER), being equal to 0.00179533, falls within the 90% central confidence interval for the determinant of the normal rank correlation matrix (DNR) [0.0014418, 0.0029218].

Under this assumption, the Bayesian network is estimated based on the model selection procedure described in the theoretical Sections, i.e., first of all, by transforming the variables into normal scores with Winsorized truncation and then running the *Hill Climbing* algorithm fixing the number of bootstrap samples at 500 and using the AIC criterion.

The threshold value is set to 0.70; in this way, only the arcs appearing at least the 70% of times are retained. This high value allows to define a parsimonious model avoiding that in the parameters' learning phase too many arcs are associated with a conditional rank correlation close to 0.[6]

We argue that the model selection procedure is mainly data-driven even if some arcs' direction are forbidden based on logical constraints.

Given the DAG structure, the conditional correlations associated with the arcs are then estimated based on the Gaussian copula assumption through the `Uninet` software.

---

[6] Moreover, this threshold provides almost the same model as the one based on the BIC criterion (with a lower threshold).

The estimated NPBN is shown in Fig. 2a. The number associated with each arc is the conditional normal rank correlation, remembering that the rank correlation between the node and its first parent is unconditional, all the others are conditioned to the previous parents' nodes. The ordering between parents is meaningful; hence, for each node, it has been set based on expert knowledge.

In the Fig. 2b, the same network is reported whose nodes are replaced by monitors showing the associated empirical marginal distributions (with the indication of the mean and the standard deviation) estimated from data.

The estimated network describes the complexity of the multivariate dependence structure discovering the conditional independence relationships among the performance indicators.

In particular, the concept of the Markov blanket can be useful to identify the main variables that influence, for example, the player's winning percentage. By definition, the Markov blanket of a node X, denoted by $MB(X)$, includes all the parents, children and parents of children of X. It contains all information needed to infer a target variable X, making all variables not belonging to it redundant. Regarding the target node "WIN", its Markov Blanket is composed of all nodes coloured blue in Fig. 2.

In terms of conditional independence, this implies that the winning percentage and the green variables are conditionally independent, given the blue ones.

This means that the player winning percentage is completely determined by the player offensive and defensive ratings (OFFRTG and DEFRTG), with nearly the same rank correla-
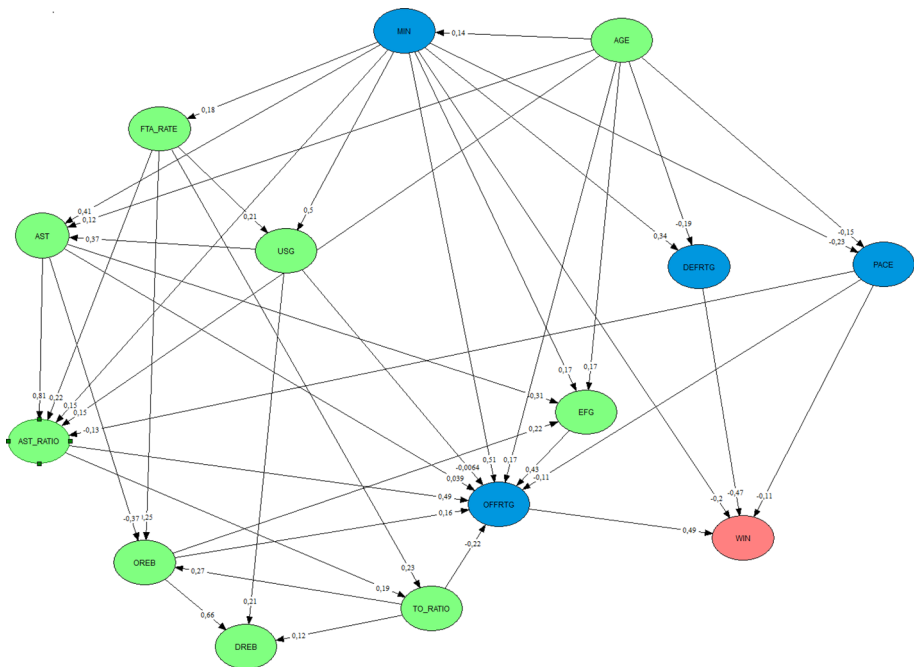
(a) The estimated network

Fig. 2 The estimated averaged BN under the assumption of Gaussian Copula, using the Uninet software representation
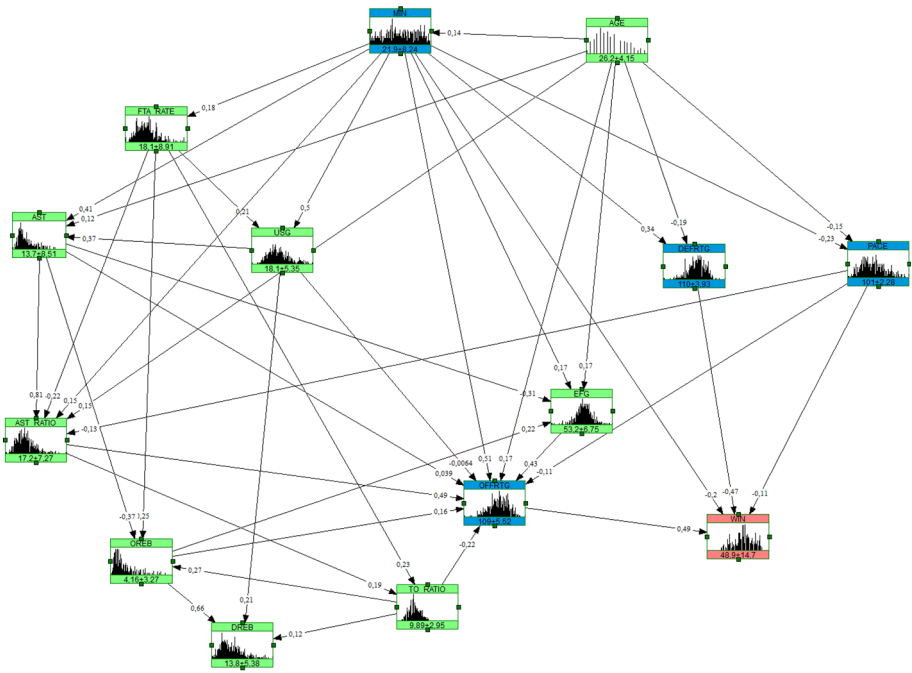
(b) The estimated network with nodes replaced by monitors showing the empirical distributions

Fig. 2 continued

tion even if with opposite sign,[7] by the possession (PACE) and by the minutes staying on the floor (MIN). No other information is needed to predict winning percentage if those related to the above variables are available.

This is in accordance with the SEM model proposed by Baghal (2012), whose two latent factors affecting winning percentage were the offensive and defensive quality factors influenced, in turn, by the Four Factors. This is also verified in the BN since the Four Factors affect the winning percentage by means of indirect paths; indeed, unless FTA_RATE, all the other three are parent nodes of the OFFRTG. Moreover, from the estimated network, the important role of the effective field percentage is confirmed as the main factor influencing the offensive quality.

In general, all the discovered conditional relationships well match with those expected or defined in the literature.

In the following, in order to use the BN for inference purposes, we simulated the following scenario focusing on the importance of the individual Offensive and Defensive Rating, i.e. the player's offensive production and the player's defensive production. Taking into account an average player, who plays 26 min and is 27 years old, the winning percentage is 49.5% (see Fig. 3a).

---

[7] We argue that the rank correlation between OFFRTG and WIN is marginal, that between DEFRTG and WIN is conditional to OFFRTG. To this purpose, changing the ordering of these two parents nodes, we observed that the marginal rank correlation between DEFRTG and WIN is equal to -0.34 while the conditional rank correlation between OFFRTG and WIN given DEFRTG, rises to 0.57.

Suppose we have an OFFRTG of 10% higher than the mean, thus setting the value of the OFFRTG node at 121 (see Fig. 3b), the mean value of the WIN node goes from 49.5 to 69.2%. Suppose now that we decrease the mean value of the DEFRTG node by 10%, fixing the value of the node at 99 (see Fig. 3c), the mean value of the WIN node rises to 61.7%. It means that a 10% increase in offensive quality implies a 40% relative increase in the winning percentage while the same variation in terms of defensive quality implies a 25% relative increase in the winning percentage. Therefore, the quality of the player offensive production seems to have a stronger impact on the winning percentage.

By looking at the monitors, the effects on the whole distribution can be seen as they allow the visualization of both the original unconditional distribution (showed in grey) and the new conditional distribution (showed in black).

In the second scenario, we vary the mean value of the Four Factors by 10%, given the same average player. This combination leads to a relative increase in the winning percentage of 7% (see Fig. 4 compared to Fig. 3a), all other factors being equal, the DEFRTG above all.

More interesting could be the third simulated scenario in which we follow a bottom-up approach. This time, we fix the expected high level of winning percentage at 75% in order to know what are the minimum levels of the other factors that guarantee to reach the above percentage (see Fig. 5 compared with Fig. 3a). Amongst them, the main evidence relates to the marked increase of Net Rating, i.e. the difference between OFFRTG and DEFRTG Obviously, the other indicators also change accordingly and can be investigated by looking at the change in their mean values (and in their distributions) compared to the baseline ones in Fig. 3a.

In the fourth scenario we investigate the role of the PACE variable. Being equal the values of OFFRTG and DEFRTG (see Fig. 6a), decreasing its value to 98 increases the mean value
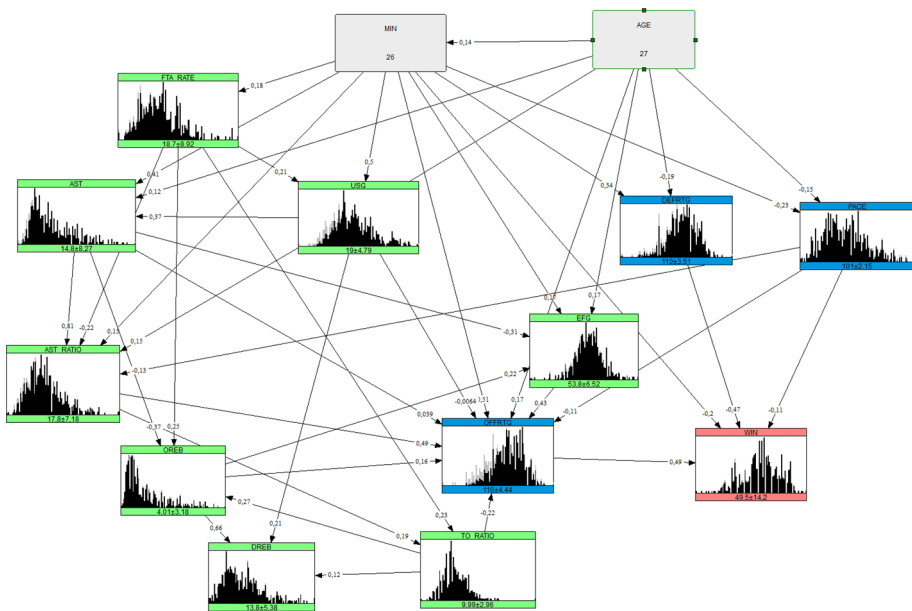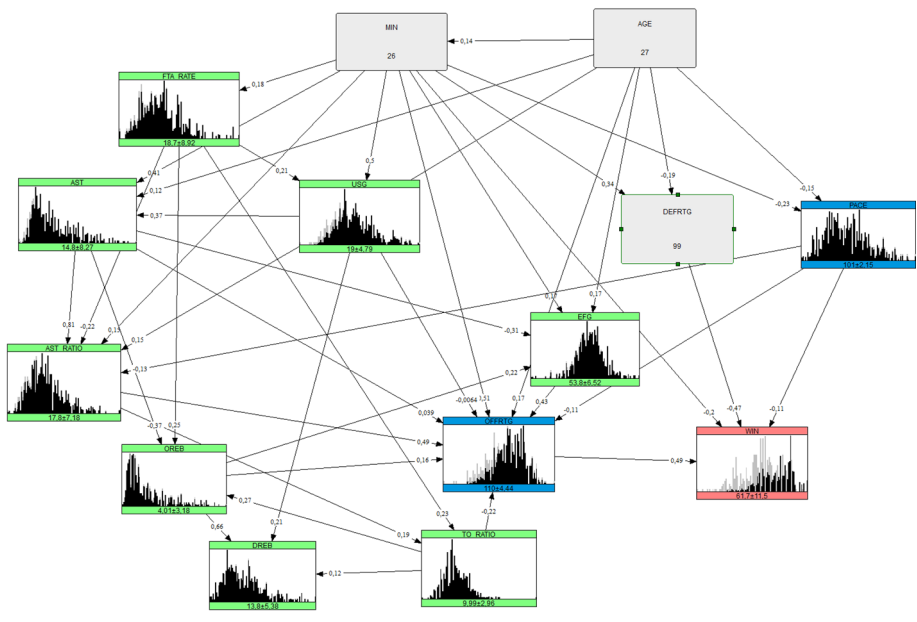

(a)

Fig. 3 First simulated scenario
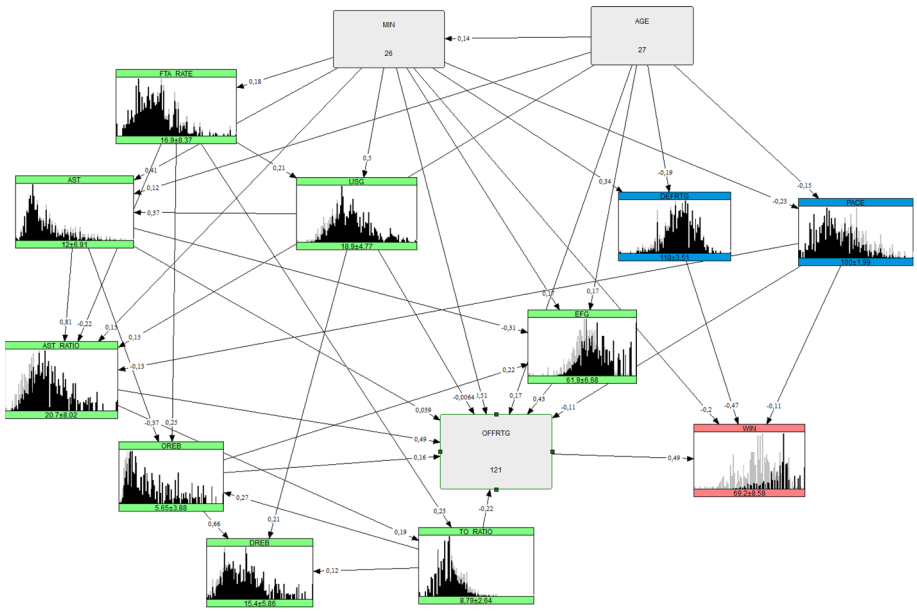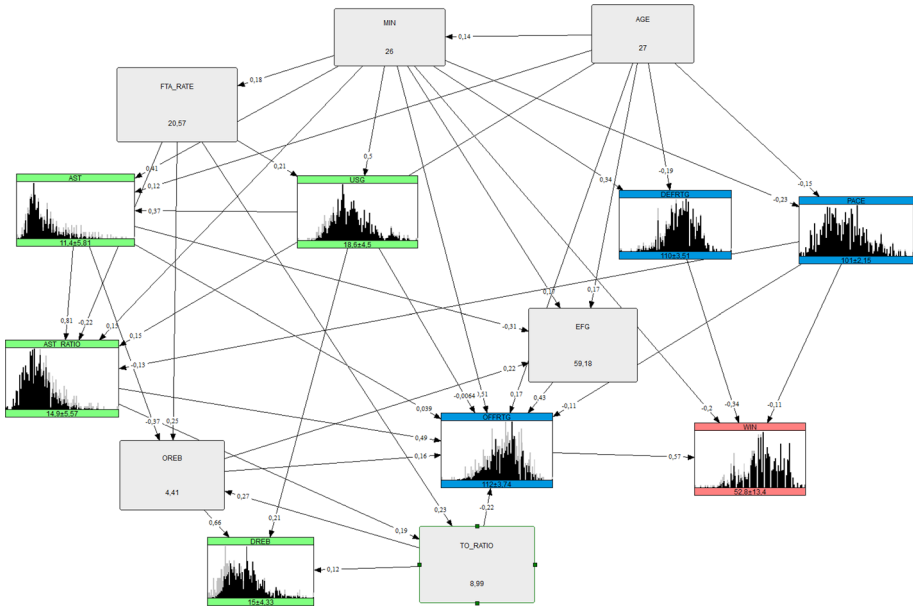
(b)

(c)

Fig. 3 continued

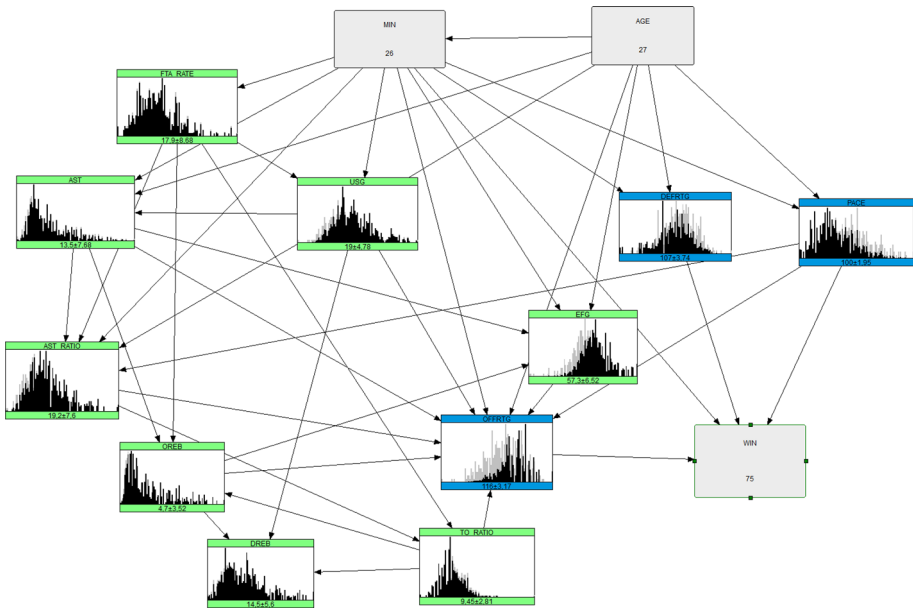**Fig. 4** Second simulated scenario

Fig. 5 Third simulated scenario

of the winning percentage from 50.8 to 52.7 (see Fig 6b) while if it grows up to 105, the mean value of the winning percentage decreases to 48.6 (see Fig 6c).

At last, to investigate the influence of the player's age, we simulate a fifth scenario. Being equal all other factors' configuration, by increasing the age from 26 (Fig. 7a) to 35 (Fig. 7b), the mean value of the winning percentage goes from 53.7 to 59.4. This could be explained by the higher value of the OFFRTG which also corresponds to an increase in the AST_RATIO for the older player.

By these simulations, we have shown the potential uses of BNs in this context, suggesting that any other scenario of interest could be investigated and deeply analyzed.

6 Conclusions

The aim of this work is to contribute to sports analytics by introducing a multivariate approach and, therefore, more complex statistical models, such as BNs, to deal with many performance indicators at the same time, studying their interrelationships rather than analysing them individually.

As shown through the *what-if* analysis, the links among different measures are evident. By means of the concept of Markov Blanket, we also identified the key determinants of the winning percentage. Moreover, the role of the Four Factors has been analysed: they affect the quality of the offensive and defensive production of a player that, in turn, directly affect the winning percentage. From a methodological point of view, the use of an approach overcoming the Gaussian assumption by means of a Gaussian Copula, is of interest in this field and, in general, in Statistics since it is well-known to allow great flexibility in the models. In the
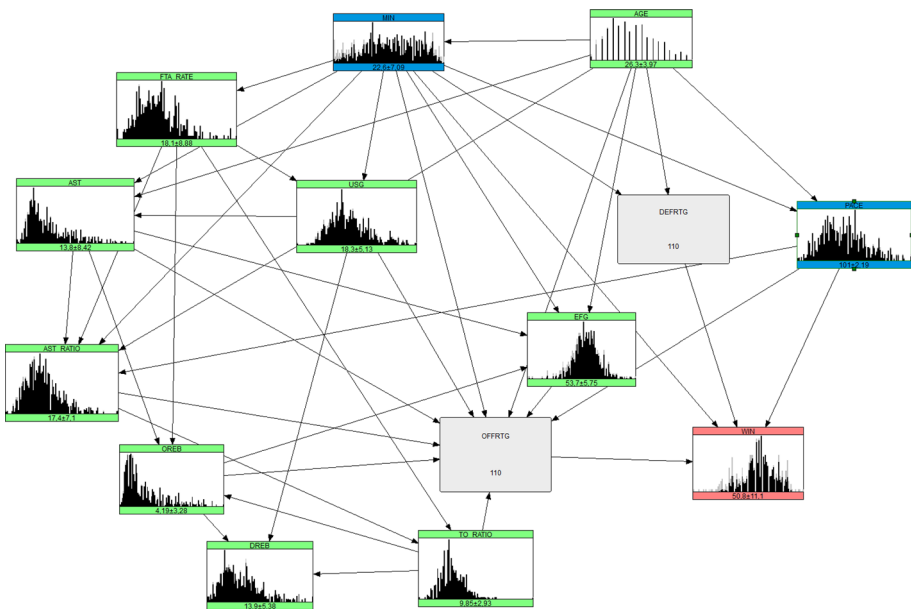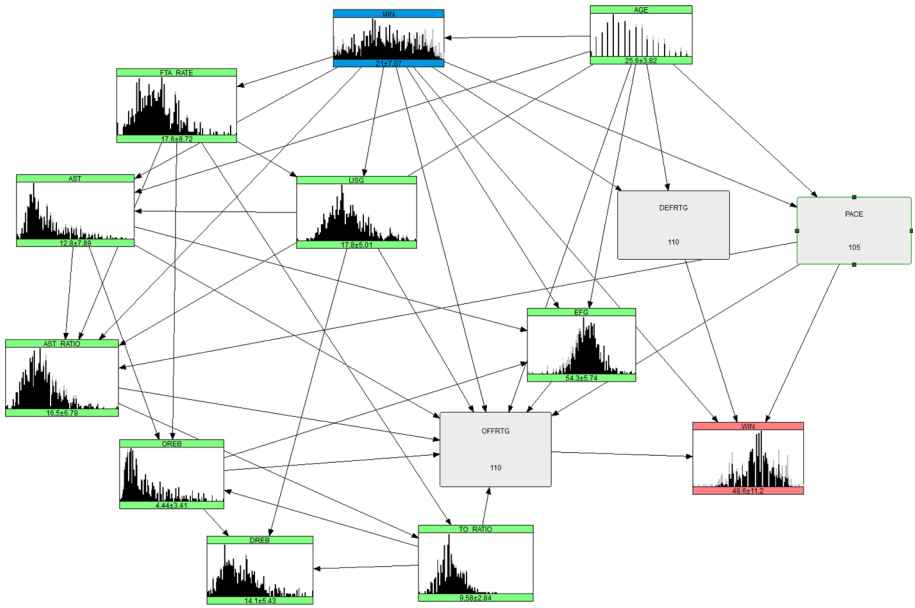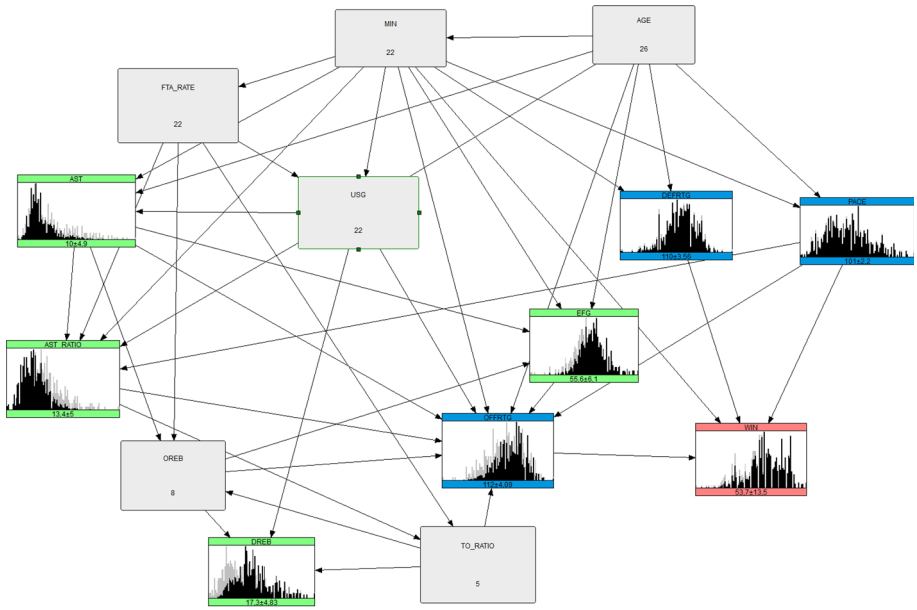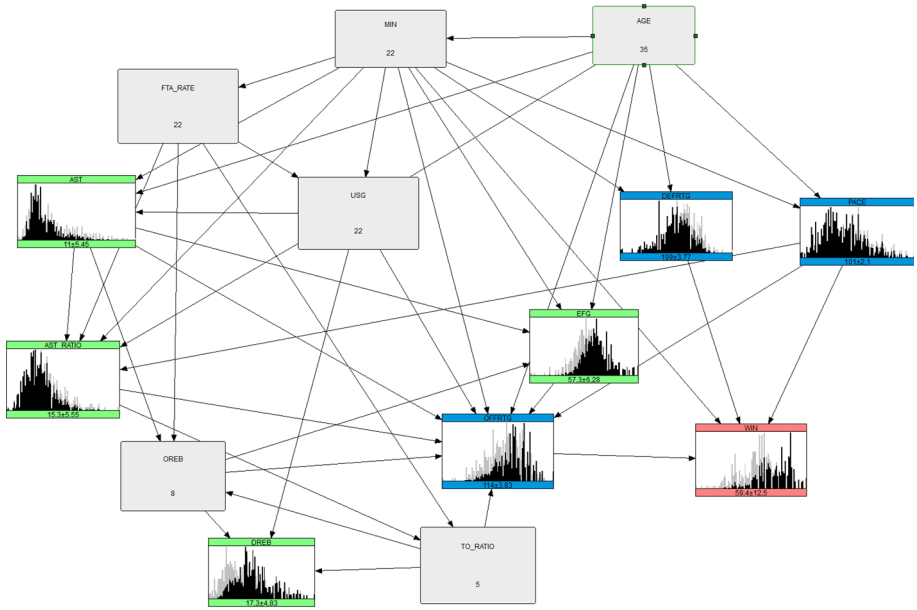

(a)

Fig. 6 Fourth simulated scenario

(b)

(c)
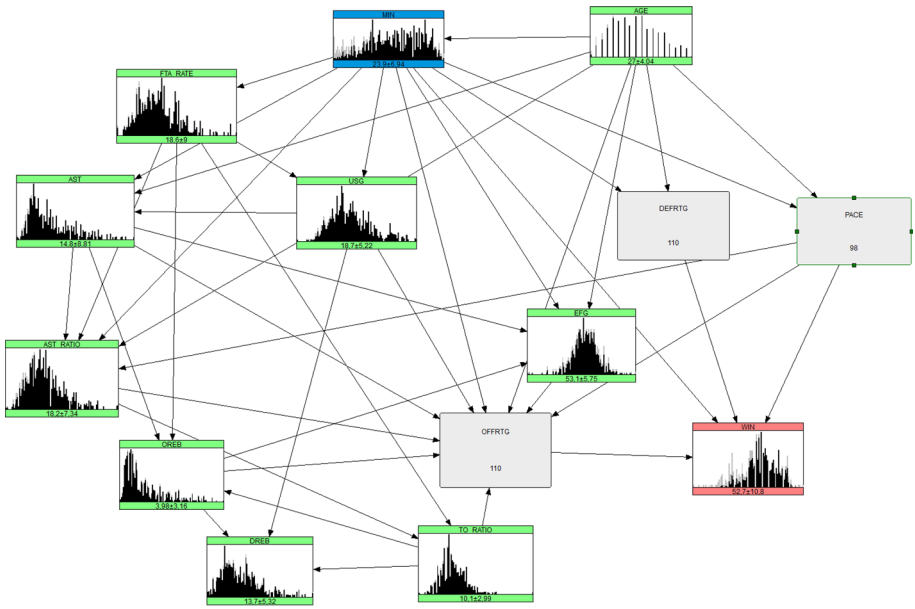
Fig. 6 continued

Fig. 7 Fifth simulated scenario

future, other statistical models based on BNs could be proposed that take into account the position of the player and many other features not yet included, such as the different seasons of the league.

References

Babaee Khobdeh, S., Yamaghani, M. R., & Khodaparast Sareshkeh, S. (2021). Clustering of basketball players using self-organizing map neural networks. *Journal of Applied Research on Industrial Engineering, 8*(4), 412–428.

Baghal, T. (2012). Are the "four factors" indicators of one factor? an application of structural equation modeling methodology to nba data in prediction of winning percentage. *Journal of Quantitative Analysis in Sports*, *8*(1).

Bauer, A., & Czado, C. (2016). Pair-copula Bayesian networks. *Journal of Computational and Graphical Statistics, 25*(4), 1248–1271.

Bauer, A., Czado, C., & Klein, T. (2012). Pair-copula constructions for non-Gaussian dag models. *Canadian Journal of Statistics, 40*(1), 86–109.

Blaikie, A. D., Abud, G. J., David, J. A., Pasteur, R. D. (2011). "nfl & ncaa football prediction using artificial neural network". In *Proceedings of the midstates conference for undergraduate research in computer science and mathematics*, Denison University, Granville, OH.

Buntine, W. (1996). A guide to the literature on learning probabilistic networks from data. *IEEE Transactions on Knowledge and Data Engineering, 8*(2), 195–210.

Casals, M., & Martinez, A. J. (2013). Modelling player performance in basketball through mixed models. *International Journal of Performance Analysis in Sport, 13*(1), 64–82.

Cervone, D., D'Amour, A., Bornn, L., & Goldsberry, K. (2016). A multiresolution stochastic process model for predicting basketball possession outcomes. *Journal of the American Statistical Association, 111*(514), 585–599.

Constantinou, A. C., Fenton, N. E., & Neil, M. (2013). Profiting from an inefficient association football gambling market: Prediction, risk and uncertainty using Bayesian networks. *Knowledge-Based Systems, 50,* 60–86.

Cooper, G. F. (1997). A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. *Data Mining and Knowledge Discovery, 1*(2), 203–224.

Cooper, G. F., & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning, 9*(4), 309–347.

Cowell, R. G., Dawid, P., Lauritzen, S. L., & Spiegelhalter, D. J. (1999). *Probabilistic networks and expert systems*. Springer.

Dalla Valle, L., & Kenett, R. S. (2015). Official statistics data integration for enhanced information quality. *Quality and Reliability Engineering International, 31*(7), 1281–1300. https://doi.org/10.1002/qre.1859 (**in Press**).

Deshpande, S. K., & Jensen, S. T. (2016). Estimating an nba player's impact on his team's chances of winning. *Journal of Quantitative Analysis in Sports, 12*(2), 51–72.

Elidan, G. (2010). Copula Bayesian networks. *Advances in Neural Information Processing Systems*, *23*, 559–567.

Engelmann, J. (2017). Possession-based player performance analysis in basketball (adjusted+/–and related concepts). In *Handbook of statistical methods and analyses in sports*. Chapman and Hall (pp. 231–244).

Fearnhead, P., & Taylor, B. M. (2011). On estimating the ability of nba players. *Journal of Quantitative analysis in sports*, *7*(3)

Fewell, J. H., Armbruster, D., Ingraham, J., Petersen, A., & Waters, J. S. (2012). Basketball teams as strategic networks. *PLoS ONE, 7*(11), e47445.

Friedman, N., Goldszmidt, M., & Wyner, A. (1999a). Data analysis with Bayesian networks: A bootstrap approach. In *Proceedings of the 15th annual conference on uncertainty in artificial intelligence* (pp. 196–201).

Friedman, N., Nachman, I., & Pe'er, D. (1999b). Learning Bayesian network structure from massive datasets: The "sparse candidate" algorithm. In M. Kaufmann (Ed.), *Proceedings of 15th conference on uncertainty in artificial intelligence* (pp. 206–221).

Geiger, D., & Heckerman, D. (1994). Learning gaussian networks. In *Proceedings of the tenth international conference on uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (pp. 235–243).

Hanea, A. M., Kurowicka, D., & Cooke, R. M. (2006). Hybrid method for quantifying and analyzing Bayesian belief nets. *Quality and Reliability Engineering International, 22*(6), 613–729.

Hanea, A., Kurowicka, D., Cooke, R., & Ababei, D. (2010). Mining and visualising ordinal data with non-parametric continuous bbns. *Computational Statistics and Data Analysis, 54*(3), 668–687.

Harris, N., & Drton, M. (2013). Pc algorithm for nonparanormal graphical models. *Journal of Machine Learning Research*, *14*(11), 3365–3383.

Heckerman, D. (1995). A tutorial on learning with Bayesian networks. In *Technical report*.

Hobæk Haff, I., Aas, K., Frigessi, A., & Lacal, V. (2016). Structure learning in Bayesian networks using regular vines. *Computational Statistics and Data Analysis, 101*(C), 186–208.

Hollinger, J. (2004). Pro basketball forecast 2004–2005. Brassey's.

Hollinger, J., & Hollinger, J. (2005). *Pro basketball forecast, 2005–2006*. Potomac Books.

James, B. (1984). *The bill James baseball abstract 1987*. Ballantine Books.

James, B. (1987). *The bill James baseball abstract 1987*. Ballantine Books.

Jensen, F. V., Lauritzen, S. L., & Olesen, K. G. (1990). Bayesian updating in causal probabilistic networks by local computations. *Computational Statistics Quarterly, 4*, 269–282.

Joe, H., & Kurowicka, D. (2011). *Dependence modeling: Vine copula handbook*. World Scientific.

Karra, K., & Mili, L. (2016). Hybrid copula Bayesian networks. In *Conference on probabilistic graphical models*. PMLR (pp. 240–251).

Kubatko, J., Oliver, D., Pelton, K., & Rosenbaum, D. T. (2007). A starting point for analyzing basketball statistics. *Journal of Quantitative Analysis in Sports*, *3*(3), 1–24.

Kurowicka, D., & Cooke, R. (2006). *Uncertainty analysis with high dimensional dependence modelling*. Wiley.

Kurowicka, D., & Cooke, R. (2010). Vines and continuous non-parametric Bayesian belief nets with emphasis on model learning, vol Re-Thinking Risk Measurement and Reporting. *Uncertainty, Bayesian Analysis and Expert Judgement, Risk Books, London, chap, 24*, 295–329.

Lauritzen, S. L., Dawid, A. P., Larsen, B. N., & Leimer, H. G. (1990). Independence properties of directed Markov fields. *Networks, 20*(5), 491–505.

Lauritzen, S. L., & Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society, Series B, 50*(2), 157–224.

Lauritzen, S. L., & Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *The annals of Statistics*, *17*(1), 31–57.

Liu, H., Han, F., Yuan, M., Lafferty, J., Wasserman, L., et al. (2012). High-dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics, 40*(4), 2293–2326.

Liu, H., Lafferty, J., & Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, *10*(10), 2295–2328.

Loeffelholz, B., Bednar, E., & Bauer, K. W. (2009). Predicting nba games using neural networks. *Journal of Quantitative Analysis in Sports*, *5*(1), 1–17.

Lopez, M. J., & Matthews, G. J. (2015). Building an ncaa men's basketball predictive model and quantifying its success. *Journal of Quantitative Analysis in Sports, 11*(1), 5–12.

Marella, D., Vicard, P., Vitale, V., & Ababei, D. (2019). Measurement error correction by nonparametric Bayesian networks: Application and evaluation. In *Statistical learning of complex data. CLADAG 2017. Studies in classification, data analysis, and knowledge organization* (pp. 155–162). Springer.

Metulini, R., Manisera, M., & Zuccolotto, P. (2018). Modelling the dynamic pattern of surface area in basketball and its effects on team performance. *Journal of Quantitative Analysis in Sports, 14*(3), 117–130.

Neapolitan, R. E. (2003). *Learning Bayesian networks*. Prentice-Hall Inc.

Nikolaidis, Y. (2015). Building a basketball game strategy through statistical analysis of data. *Annals of Operations Research, 227*(1), 137–159.

Nojavan, F., Qian, S. S., & Stow, C. A. (2017). Comparative analysis of discretization methods in Bayesian networks. *Environmental Modelling & Software, 87,* 64–71.

Oliver, D. (2004). *Basketball on paper: Rules and tools for performance analysis*. Potomac Books, Inc.

Page, G. L., Fellingham, G. W., & Reese, C. S. (2007). Using box-scores to determine a position's contribution to winning basketball games. *Journal of Quantitative Analysis in Sports*, *3*(4), 1–18.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann Publishers Inc.

Piette, J., Pham, L., & Anand, S. (2011). Evaluating basketball player performance via statistical network modeling. In *The 5th MIT Sloan sports analytics conference*.

Pircalabelu, E., Claeskens, G., & Gijbels, I. (2017). Copula directed acyclic graphs. *Statistics and Computing, 27*(1), 55–78.

Razali, N., Mustapha, A., Yatim, F. A., & Ab Aziz, R. (2017). Predicting football matches results using Bayesian networks for English premier league (epl). In *Iop conference series: Materials science and engineering* (Vol. 226, p. 012099). IOP Publishing.

Rohmer, J. (2020). Uncertainties in conditional probability tables of discrete Bayesian belief networks: A comprehensive review. *Engineering Applications of Artificial Intelligence, 88,* 103384.

Sandholtz, N., Mortensen, J., & Bornn, L. (2020). Measuring spatial allocative efficiency in basketball. *Journal of Quantitative Analysis in Sports, 16*(4), 271–289.

Scutari, M., & Nagarajan, R. (2011). On identifying significant edges in graphical models. In *Proceedings of workshop on probabilistic problem solving in biomedicine* (pp. 15–27). Springer.

Shen, J., Zhao, Y., Liu, J. K., & Wang, Y. (2020). Recognizing scoring in basketball game from AER sequence by spiking neural networks. In *2020 international joint conference on neural networks (IJCNN)* (pp. 1–8). IEEE.

Sill J (2010) Improved nba adjusted+/-using regularization and out-of-sample testing. In *Proceedings of the 2010 MIT Sloan sports analytics conference*.

Skinner, B., & Guy, S. J. (2015). A method for using player tracking data in basketball to learn player skills and predict team performance. *PLoS ONE, 10*(9), e0136393.

Spirtes, P., Glymour, C., & Scheines, R. (1993). Discovery algorithms for causally sufficient structures. In *Causation, prediction, and search* (pp. 103–162). Springer.

Terner, Z., & Franks, A. (2021). Modeling player and team performance in basketball. *Annual Review of Statistics and Its Application, 8,* 1–23.

Tsamardinos, I., Brown, L. E., & Aliferis, C. F. (2006). The max–min hill-climbing Bayesian network structure learning algorithm. *Machine Learning, 65*(1), 31–78.

Vitale, V., Musella, F., Vicard, P., & Guizzi, V. (2018) Modelling an energy market with Bayesian networks for non-normal data. *Computational Management Science*, 1–18.

Wang, K. C., & Zemel, R. (2016). Classifying nba offensive plays using neural networks. In: *Proceedings of MIT Sloan sports analytics conference* (Vol. 4).

Wu, S., & Bornn, L. (2018). Modeling offensive player movement in professional basketball. *The American Statistician, 72*(1), 72–79.

Xin, L., Zhu, M., & Chipman, H. (2017). A continuous-time stochastic block model for basketball networks. *The Annals of Applied Statistics, 11*(2), 553–597.

Yang, C. H., Lin, H. Y., & Chen, C. P. (2014). Measuring the efficiency of nba teams: Additive efficiency decomposition in two-stage dea. *Annals of Operations Research, 217*(1), 565–589.

Zuccolotto, P., Sandri, M., & Manisera, M. (2019). Spatial performance indicators and graphs in basketball. *Social Indicators Research*, 1–14.