Copula Gaussian graphical models with penalized ascent Monte Carlo EM algorithm

Fentaw Abegaz* and Ernst Wit†

Johann Bernoulli Institute of Mathematics and Computer Science,
University of Groningen, Nijenborgh 9, 9747 AG Groningen,
The Netherlands

Typical data that arise from surveys, experiments, and observational studies include continuous and discrete variables. In this article, we study the interdependence among a mixed (continuous, count, ordered categorical, and binary) set of variables via graphical models. We propose an $\ell_1$-penalized extended rank likelihood with an ascent Monte Carlo expectation maximization approach for the copula Gaussian graphical models and establish near conditional independence relations and zero elements of a precision matrix. In particular, we focus on high-dimensional inference where the number of observations are in the same order or less than the number of variables under consideration. To illustrate how to infer networks for mixed variables through conditional independence, we consider two datasets: one in the area of sports and the other concerning breast cancer.

Keywords and Phrases: Gaussian copula, $\ell_1$-penalized maximum likelihood, Gaussian graphical models, ascent-MCEM algorithm, extended rank likelihood, conditional independence.

1 Introduction

Data that arise from surveys, experiments, and observational studies include typically continuous and discrete variables. The study of dependence pattern among mixed variables that involve continuous, binary, and ordered categorical types requires the determination of their joint distribution. The copula framework allows the joint distribution to be represented by univariate marginal distributions and a copula function. As a result, multivariate dependence, which is fully described by the copula function, can be modeled separately from the univariate marginal distributions. Many types of two dimensional copulas exist (Joe, 1997; Nelsen, 2006); however, their extensions to multivariate copulas have been limited. So far, the Gaussian copula has been the natural choice beyond the bivariate case.

The Gaussian copula has been considered for the study of dependence pattern among mixed variables through the extended rank likelihood estimation under a

*abegaz.yazew@rug.nl
†e.c.wit@rug.nl

© 2015 The Author. Statistica Neerlandica © 2015 VVS.
Published by Wiley Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.
Bayesian framework by Hoff (2007) and further studied in the graphical model setting by Dobra and Lenkoski (2011) using the Bayesian model averaging approach for graph identification and estimation. In the extended rank likelihood approach using the Gaussian copula, because the marginals are treated as nuisance parameters, the parameter of interest for estimation is the correlation matrix or the precision matrix, that is, the inverse of the correlation matrix in the case of a Gaussian copula. Ambroise et al. (2009) raised their concern on the challenging task involved in the Bayesian framework to construct priors on the set of precision matrices. We propose an alternative approach to consider the extended rank likelihood under the maximum likelihood setting with an ascent Monte Carlo expectation–maximization (EM) algorithm. Moreover, we extend this approach to a high-dimensional inference based on graphical models using $\ell_1$-penalized maximum likelihood (Banerjee et al., 2008; Friedman et al., 2008; Rothman et al., 2008; Yuan and Lin, 2007).

In graphical models, conditional independence is a fundamental concept. Conditional independence relations are encoded in the graph by Markov properties, which result in a parsimonious representation of probabilistic models and a sparse graphical representation convenient for describing complex interdependence among a set of variables. In the case that all variables are continuous and assumed to follow a Gaussian or Gaussian copula-based distribution, zero elements of the precision matrix encode conditional independence between two variables given the rest. Similarly, conditional independence is represented by zero coefficients in the Ising models for binary variables and in log-linear models for discrete variables in general. Local log-odds ratios are also used to describe partial association between two discrete variables at each level of a third conditioning variable (Becker, 1989). In these approaches, when the conditioning variables are discrete, conditional (in)dependence or partial association is defined for each level of the conditioning discrete variables separately. This brings an enormous increase in the set of parameters to be estimated with an increase in the number of discrete variables.

The case of mixed continuous and ordered discrete variables where their underlying process is assumed to follow a continuous distribution, assessing conditional independence of the observed variables, requires special attention. For example, Cox and Wermuth (1992, 1999) addressed properties of conditional independence for mixed variables from a partially dichotomized Gaussian distribution. They noted that the only situation that retains conditional independence of the underlying Gaussian distribution is when all involved conditioning variables are continuous. In all other cases, strong additional independence statement is required to hold. Recently, Murray et al. (2013) showed that for the copula Gaussian distributions, zeros in the precision matrix generally do not imply conditional independence when some variables are discrete. In such a case, they noted that the interpretation of conditional independence requires care. Similarly, García-Zattera et al. (2007) showed that for binary multivariate variables, conditional independence on the latent scale does not imply conditional independence on the observed scale. We note, however, that in these studies, the role played by increasing the number of categories or levels in ordered discrete data has been
overlooked. In this work, we investigate the possibility of retaining conditional independence in the Gaussian copula graphical models for mixed continuous and ordered discrete variables.

Additional advantage of the proposed Gaussian copula graphical modeling for studying the relationship among a set of mixed types of continuous and discrete variables with an ascent-based $\ell_1$-penalized EM algorithm is to handle missing data in the E-step of the EM algorithm without involving extra computation. To illustrate the proposed approach, we consider two high-dimensional mixed continuous and ordered discrete datasets on sports and breast cancer. The breast cancer dataset includes missing values.

Sports play an enormous role as a medium for physical and socio-economic development, to bridge cultural gaps and to contribute to social integration and tolerance. Successful sport performance can be achieved possibly by combining expert knowledge and relevant information extracted from the data. With the increase of automated sampling devices, a huge amount of sports data is collected for each player, team, game, and season. Mostly, such sports data consist of many variables of mixed type measured on a comparatively small number of teams and players. As a result, high-dimensional statistical techniques for mixed data are required for a proper analysis of sports data to support decision making.

As an example of high-dimensional sports data, we focus on the American football, which is one of the most popular sports in the USA and has a following around the world. In Netherlands, although the American football is not so popular, for example, in 2014, the American Football Bond Nederland Premier League includes six teams split into two divisions. Because of the availability of data, we consider the American National Football League (NFL) in the USA that comprises 32 teams, organized into two conferences: the American Football Conference and the National Football Conference, each containing 16 teams. For many stakeholders, such as players, coaches, owners, fans, and advertisers, an important aspect of the American football game is to identify what contributes to the outcome of an NFL match. There are limited scholarly research works that examine the correlates of game performances and game outcomes that mainly use logistic or probit regression and factor analysis or principal component approaches (Watnik and Levine, 2001; White and Berry, 2002; Boulier and Stekler, 2003). Recently, Baker and McHale (2013) considered a point process model for predicting the exact end-of-match scores in the NFL.

Data from the NFL football games cover a vast amount of information, such as home versus away, yards gained and allowed, passing, rushing, punts, interceptions, fumbles, and points. Factors and variables that are considered important to affect game outcome can be grouped into four headings: general factors that include total yards, time of possession, and home field advantage; offensive factors that include third-down, fourth-down, first-down, and second-down conversions, passing and rushing efficiencies, red zone conversions, penalties, and penalty yards; defensive factors that include sacks, tackling efficiency, interceptions, and red zone stops; and special teams factors that include kick-offs and kick returns, punts and punt returns, and field
goals (Cohea and Payton, 2011). Our interest is to understand how these variables are interrelated to each other from the point of view of offense–defense strategy and game outcome variables.

We also consider a second application of our approach on breast cancer genomic and clinical data. Breast cancer is the leading cause of death among women in the world and represents a significant health problem. Multiple factors like age, diet, obesity, parity, age at first childbirth, oral contraceptives, exogenous estrogens, genetics, environment, and geographic location influence the development of breast cancer. However, the majority of breast cancer cases are always due to genetic abnormalities (Kumar et al., 2012). At present, only small numbers of accurate prognostic and predictive factors are used clinically for managing the patients with breast cancer. In the last few decades, knowledge of breast cancer grade determined by the Nottingham prognostic index (NPI) has been very helpful to decide on the most effective treatments. Moreover, microarray-based gene expression profiling has been used extensively to characterize the transcriptome of breast cancer, resulting in the identification of new molecular subtypes and markers or signatures of potential therapeutic and prognostic importance (Ringnér et al., 2011). Inclusion of such treatment predictive markers considerably improved breast cancer treatment decisions. To further tailor treatment for individual patients, identification of additional clinical and genetic markers is required.

Genomic DNA copy number alterations, that is, amplifications or deletions of genetic materials, are key genetic events in the development and progression of breast cancers. Gene copy number changes can be determined on a gene-by-gene basis using microarrays. A genome-wide microarray comparative genomic hybridization is used to analyze the pattern of DNA copy number alteration with the aim to study the relationship between DNA amplification and deletion patterns and severity of breast cancer as measured by several clinical indicators on patients. Details of the experimental setup is discussed by Wit and McClure (2004) and Witton et al. (2002). The breast cancer data are mixed measurements of continuous, binary, and ordered categorical types; see details in Section 4.

The article is organized as follows. In section 2, we briefly discuss the Gaussian copula graphical modeling aspects related to continuous, binary, count, and ordinal variables. Moreover, we discuss conditional independence relationship between ordered discrete variables and the underlying continuous distribution. In section 3, we present the EM-based $\ell_1$-penalized maximum likelihood estimation using the Gaussian copula. We also present a method of model selection. In section 4, we analyze two datasets and discuss the results. The last section presents a discussion.

2 Graphical models

Graphical models are efficient tools for studying multivariate distributions through a compact, graphical representation of the joint probability distribution of the underlying random variables. Consider an undirected graph $G = (V, E)$, where $V = \{1, \ldots, p\}$ corresponds to the set of nodes or vertices of the graph $G$ with $p$ elements and $E \subset V \times V$
of ordered pairs of distinct nodes called the edges of $G$. If we associate the nodes of the graph with the random variables $Y = (Y_1, \ldots, Y_p)$, then, roughly, the joint distribution of $Y$ is called a graphical model with respect to graph $G$, if the conditional independence relationships between the variables $Y_1, \ldots, Y_p$ are described by the absence of the associated links in $E$. For a formal definition of a graphical model, the reader is referred to Lauritzen (1996).

2.1 Gaussian graphical models

The treatment of graphical models simplifies significantly, when one focuses on normally distributed variables. Let the random vector $Y = (Y_1, \ldots, Y_p)^T$ be the Gaussian with a positive definite covariance matrix $\Sigma$ of the dimension $p \times p$. Without loss of generality, we assume $Y$ follows a $p$-dimensional multivariate normal distribution with mean zero and covariance matrix $\Sigma$, $N(0, \Sigma)$. A graphical model $G = (V, E)$ for $N(0, \Sigma)$ is called a Gaussian graphical model, if on the graph $G$, the edges $E$ represent conditional dependence among the random variables. The absence of an edge between a pair of entries $Y_j$ and $Y_k$ corresponds to the conditional independence of these two random variables given the remaining variables $Y_{p \setminus j,k}$, where the index $p \setminus j,k$ refers to variables other than those indexed by $j$ and $k$.

The precision matrix, also known as the concentration matrix, is the inverse of the covariance matrix, $\Theta = \Sigma^{-1}$. A standard result (Whittaker, 1990) relates each entry of the precision matrix $\Theta_{jk}, j \neq k$ to the partial correlation coefficient $\rho_{jk \mid p \setminus j,k}$ between variables $Y_j$ and $Y_k$ via

$$
\rho_{jk \mid p \setminus j,k} = -\frac{\Theta_{jk}}{\sqrt{\Theta_{jj}\Theta_{kk}}}
$$

so that $\rho_{jk \mid p \setminus j,k} = 0 \Leftrightarrow \Theta_{jk} = 0$. Thus, under the Gaussian graphical models, the criteria to assess conditional independence of random variables is based on zero entries of the precision matrix.

2.2 Gaussian copula graphical models

In practice, we encounter both discrete and continuous variables that may not be the Gaussian. Thus, the assumption of multivariate normal distribution would be too restrictive. To relax the normality requirement, we use the copula framework to construct multivariate distributions for arbitrary marginals. Because our aim is to explore relationship among many variables, we consider the Gaussian copula that shares important features of multivariate normal distribution such as a well-structured correlation matrix and assess conditional independence through zero entries of the precision matrix. As a result, the Gaussian copula-based distributions are useful for graphical modeling.

The Gaussian copula with correlation matrix $\Gamma$ of dimension $p \times p$ having $p(p - 1)/2$ free parameters is given by

$$
C(u_1, \ldots, u_p \mid \Gamma) = \Phi_p(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_p) \mid \Gamma), \quad (u_1, \ldots, u_p) \in [0, 1]^p.
$$
Let $Y = (Y_1, \ldots, Y_p)$ represent a vector of mixed variables with marginal distributions $F_j(Y_j)$, $j = 1, \ldots, p$. The Gaussian copula-based distribution function of $Y$ is given by

$$H(y | \Gamma, F_1, \ldots, F_p) = \Phi_p(\Phi^{-1}(F_1(y_1)), \ldots, \Phi^{-1}(F_p(y_p)) | \Gamma). \quad (2.1)$$

Here, $\Phi(\cdot)$ represents the cumulative distribution function (CDF) of the standard normal distribution, and $\Phi_p(\cdot | \Gamma)$ is the CDF of multivariate normal distribution $\mathcal{N}_p(0, \Gamma)$.

We aim to construct a graphical model for observed variables $Y$ of mixed type or non-Gaussian, that is, in the case that they represent a collection of continuous, binary, ordinal, or count variables. Suppose the $j$th variable $Y_j$ has univariate distribution $F_j$ with its pseudo-inverse $F_j^{-1}$ (Nelsen, 2006). In semiparametric copula modeling, the marginals are treated as nuisance parameters and estimated non-parametrically, mainly using the rescaled empirical distribution: $F_{nj}(y) = \frac{1}{n+1} \sum_{i=1}^n I\{Y_{ji} \leq y\}$, $j = 1, \ldots, p$, where $I\{\cdot\}$ is the indicator function. A copula graphical model can be constructed by introducing a vector of latent variables $Z$ as $Y_j = F_j^{-1}(\Phi(Z_j))$, $j = 1, \ldots, p$, such that $Y$ has a Gaussian copula-based distribution of the form in Equation 2.1. In the case of mixed variables, the graphical structure, that is, the conditional independence implied by the graph structure, holds exclusively on the latent variables. Implication of this assumption is discussed in Section 2.3.

Our aim is to infer the graphical structure $G$ defined by the latent variable $Z$. To relate the observed and latent variables, Hoff (2007) argued that although the $Z$s are not observable, the observed $Y$s do provide a limited amount of information about them. For the $j$th variable, because the transformation $y_{ij} = F_j^{-1}(\Phi(z_{ij}))$ is non-decreasing, observing $y_{ij} < y_{ij}'$ implies that $z_{ij} < z_{ij}'$. More generally, observing $y_j = (y_{1j}, \ldots, y_{nj})$ imposes constraints on $z_j = (z_{1j}, \ldots, z_{nj})$ to lie on the set

$$D(y_j) = \{z_j \in \mathbb{R}^n : L_{ij}(y_{ij}) < z_{ij} < U_{ij}(y_{ij})\}, \quad (2.2)$$

where $L_{ij}(y_{ij}) = \max\{z_{kj} : y_{kj} < y_{ij}\}$ and $U_{ij}(y_{ij}) = \min\{z_{kj} : y_{ij} < y_{kj}\}$.

Let the set $D_p(y) = \{z_j \in \mathbb{R}^{np} : z_j \in D(y_j), j = 1, \ldots, p\}$ denote the collection of these intervals. Because the event $Z \in D_p(Y)$ occurs whenever $Y$ is observed, the probability of the observed data can be decomposed as

$$P(Y | \Gamma, F_1, \ldots, F_p) = P(Y, Z \in D_p(Y) | \Gamma, F_1, \ldots, F_p) = P(Z \in D_p(Y) | \Gamma) P(Y | Z \in D_p(Y), \Gamma, F_1, \ldots, F_p). \quad (2.3)$$

Similar to the idea of marginal likelihood, Hoff (2007) defined the extended rank likelihood via

$$P(Z \in D_p(Y) | \Gamma) = \int_{Z \in D_p(Y)} P(Z | \Gamma) dZ. \quad (2.4)$$
which is part of the observed data likelihood that depends on $\Gamma$ in Equation 2.3. The occurrence of event $Z \in D_p(Y)$ is considered as the data and used to estimate the correlation parameter $\Gamma$ separately from the marginal distributions (Hoff, 2007; Dobra and Lenkoski, 2011; Murray et al., 2013).

2.3 Conditional independence in the copula Gaussian graphical models

In this section, we investigate to what extent conditional independence for mixed variables holds in the case of zeros in the precision matrix for a Gaussian copula graphical model by making use of an approximate relationship between the local conditional dependence function proposed by Wang (1987) and the conditional local log-odds ratios proposed by Clogg (1982). The idea is based on observing the relationship between the latent and the ordered discrete variables as a result of partitioning the multivariate normal distribution into ordered categories using rectangular grids specified by $D_j = D(y_j)$, $j = 1, \ldots, p$, in Equation 2.2. Let $c_j$ be the number of distinct values of $Y_j$ and $I(m)$ be the $m$th interval in $D_j$ that partitions $Z_j$. Then $D_1 \times D_2 \times \cdots \times D_p$ partitions the $p$-dimensional space into $c_1 \times c_2 \times \cdots \times c_p$ rectangular cubes or cells $E(i_1, \ldots, i_p) = I_1(i_1) \times \cdots \times I_p(i_p)$. So the multivariate normal probability integral over a cell $E(i_1, \ldots, i_p)$ is given by

$$p_{i_1, \ldots, i_p} = \int_{I_1(i_1)} \cdots \int_{I_p(i_p)} \phi_p(z_1, \ldots, z_p | \Theta^{-1})dz_1 \cdots dz_p$$

$$\cong \phi_p(z_{1i_1}^*, \ldots, z_{pi_p}^* | \Theta^{-1})\Delta z_{1i_1} \cdots \Delta z_{pi_p},$$

where the approximation comes from using multivariate mean value theorem, $\phi_p(z_1, \ldots, z_p | \Theta^{-1})$ is the joint multivariate normal density of $(z_1, \ldots, z_p)$, and $(z_{1i_1}^*, \ldots, z_{pi_p}^*)$ is an interior point of the $p$ dimensional cube $E(i_1, \ldots, i_p)$.

Wang (1987) provides a good approximation of the multivariate normal probability integral using a contingency table approach, which motivates us to define near conditional independence among mixed ordered variables having an underlying multivariate normal distribution. As can be seen in the following, the contingency table approach utilizes the special intrinsic dependence pattern of the multivariate normal distribution. Wang (1987) introduces the conditional dependence function, which is given by

$$\gamma_f(z_j, z_k | z_{p \setminus j, k}) = \frac{\partial^2}{\partial z_j \partial z_k} \ln f(z_j, z_k | z_{p \setminus j, k})$$

$$= \frac{\partial^2}{\partial z_j \partial z_k} \ln f(z_j, z_k, z_{p \setminus j, k}),$$

and a three-way (unconditional) dependence function given by

$$\gamma_f(z_j, z_k, z_l) = \frac{\partial^3}{\partial z_j \partial z_k \partial z_l} \ln f(z_1, \ldots, z_p).$$

The functional form of $\gamma_f(z_j, z_k | z_{p \setminus j, k})$ or $\gamma_f(z_j, z_k, z_l)$ gives a good indication of the association pattern of the discretized $Z$. In particular, $\gamma_f(z_j, z_k | z_{p \setminus j, k}) = 0$ implies that
(\(Z_j, Z_k\)) are conditionally independent given \(Z_{p\setminus j,k}\), and \(\gamma_f(z_j, z_k, z_l) = 0\) implies that there is no three-way dependence among \((Z_j, Z_k, Z_l)\).

For a multivariate normal distribution, \(f(\cdot) = \phi_p(\cdot)\), with the logarithm of the multivariate normal density factorized as (Lauritzen, 1996),

\[
\ln \phi_p(z_1, \ldots, z_p | \Theta^{-1}) = \text{constant} - \frac{1}{2} \sum_{j=1}^p \Theta_{jj} z_j^2 - \sum_{j=1}^p \sum_{j\neq k}^{p} \Theta_{jk} z_j z_k,
\]

the conditional dependence function simplifies to

\[
\gamma_f(z_j, z_k | z_{p\setminus j,k}) = \frac{\partial^2}{\partial z_j \partial z_k} \ln \phi_p(z_1, \ldots, z_p | \Theta^{-1}) = -\Theta_{jk}, \quad \text{for } j \neq k,
\]

and the three-way dependence function becomes

\[
\gamma_f(z_j, z_k, z_l) = \frac{\partial^3}{\partial z_j \partial z_k \partial z_l} \ln \phi_p(z_1, \ldots, z_p | \Theta^{-1}) = 0.
\]

Moreover, Wang (1987) established a relationship between the conditional dependence function and the discrete conditional local log-odds ratio proposed by Clogg (1982). The conditional local log-odds ratio measures the association between ordered discrete variables \(Y_j\) and \(Y_k\) conditioned on the remaining variables. It is given by

\[
\delta_{jk|\wedge(jk)} = \frac{p_{jm,k,\wedge(jk)} p_{j_{m+1},k,\wedge(jk)}}{p_{j_{m+1},k,\wedge(jk)} p_{jm,k,\wedge(jk)}}, \quad (2.5)
\]

where \(\wedge(jk)\) represents one category from each of the conditioning variables other than variables \(Y_j\) and \(Y_k\), and \(m\) and \(t\) stands for the \(m\)th and \(t\)th intervals of \(D_j\) and \(D_k\), respectively.

Applying the multivariate mean value theorem to approximate the probabilities in Equation 2.5, the relationship between \(\delta_{jk|\wedge(jk)}\) and the conditional dependence function is given by

\[
\ln \delta_{jk|\wedge(jk)} \cong \int_{z_{jm}}^{z_{jm+1}} \int_{z_{kt}}^{z_{kt+1}} \frac{\partial^2}{\partial z_j \partial z_k} \ln f(z_j, z_k, z_{\wedge(jk)}) \, dz_j dz_k,
\]

where \(\left( z_{\wedge(jk)} \right)\) is an interior point of the \(p-2\) dimensional cube after removing variables \(Z_j\) and \(Z_k\). An equivalent limiting equation is also given by

\[
\lim_{\left( \frac{\Delta z_{jm} + \Delta z_{jm+1}}{\Delta z_{jm}} \right) \to 0 \frac{\Delta z_{kt} + \Delta z_{kt+1}}{\Delta z_{kt}} \to 0} \left( \frac{4}{\Delta z_{jm} + \Delta z_{jm+1}} \Delta z_{kt} + \Delta z_{kt+1} \right) \ln \delta_{jk|\wedge(jk)}
= \frac{\partial^2}{\partial z_j \partial z_k} \ln f(z_j, z_k, z_{\wedge(jk)}) \bigg|_{z_j = d_{jm}, \ z_k = d_{kt}}.
\]
Similarly, a three-factor local log-odds ratio, \( \ln \delta_{jkl} \), is related to the local dependence function by

\[
\ln \delta_{jkl} \cong \int_{z_{jm}^*}^{z_{jm+1}} \int_{z_{km}^*}^{z_{km+1}} \int_{z_{kl}^*}^{z_{kl+1}} \frac{\partial^3}{\partial z_j \partial z_k \partial z_l} \ln f(z_j, z_k, z_l, z_{jm}) \ dz_j \ dz_k \ dz_l.
\]

For the multivariate normal latent distribution, we have

\[
\ln \delta_{jk} \mid \wedge(jk) \cong \int_{z_{jm}^*}^{z_{jm+1}} \int_{z_{km}^*}^{z_{km+1}} \frac{\partial^2}{\partial z_j \partial z_k} \ln \phi_p(z_1, z_2, \ldots, z_p | \Theta^{-1}) \ dz_j \ dz_k.
\]

The equivalent limiting equation based on the multivariate normal distribution is

\[
\lim \left( \frac{\Delta z_{jm}}{\Delta z_{jm+1}} \right) \to 0 \left( \frac{\Delta z_{km}}{\Delta z_{km+1}} \right) \to 0 \ln \delta_{jk} \mid \wedge(jk) = -\Theta_{jk}
\]

so that

\[
\ln \delta_{jk} \mid \wedge(jk) \cong -\frac{1}{4} \Theta_{jk} \left( \Delta z_{jm} + \Delta z_{jm+1} \right) \left( \Delta z_{km} + \Delta z_{km+1} \right),
\]  

(2.6)

where the accuracy of this approximation improves as \( (\Delta z_{jm} + \Delta z_{jm+1}) \) and \( (\Delta z_{km} + \Delta z_{km+1}) \) approach to zero. If \( \Theta_{jk} = 0 \), it follows that

\[
\ln \delta_{jk} \mid \wedge(jk) = o \left( \left( \Delta z_{jm} + \Delta z_{jm+1} \right) \left( \Delta z_{km} + \Delta z_{km+1} \right) \right).
\]

Because the third and higher-order partial derivatives of a multivariate normal distribution are zero, the unconditional three-factor local log-odds ratio becomes \( \ln \delta_{mkl} \cong 0 \). Becker (1989) used the difference between two consecutive conditional local log-odds ratios to define the three-way local log-odds ratios, for example, \( \ln \delta_{jkl} = \ln \delta_{jkl+1} - \ln \delta_{jkl} \cong 0 \), that implies \( \ln \delta_{jkl+1} \cong \ln \delta_{jkl} \) for all \( l \). In other words, the conditional local odds ratios of \( Y_j \) and \( Y_k \) are taken to be equal for all levels of the conditioning variable \( Y_l \).

Noting that \( \ln \delta_{jk} \mid \wedge(jk) = 0 \) is a measure of conditional independence of ordered discrete variables, we argue that conditional independence on the latent scale, that is, \( \Theta_{jk} = 0 \), results in \( \ln \delta_{jk} \mid \wedge(jk) \cong 0 \) and therefore implies ‘near’ conditional independence on the ordered discrete scale. The near conditional independence of the discrete variables tends to retain conditional independence of the underlying latent variables, when the accuracy of the approximation improves as the partitioning grids become fine enough. In practice, this could happen when the set of mixed variables involves ordered categorical variables with many categories (preferably \( \geq 5 \)), counts and continuous variables. On the other hand, the use of the Gaussian copula imposes restrictions on the dependence pattern of the discrete ordered variables such as zero higher-order interactions and same pair-wise conditional (in)dependence at all categories of ordered discrete variables.

© 2015 The Author. Statistica Neerlandica © 2015 VVS.
3 Penalized ascent Monte Carlo EM estimation

3.1 The $\ell_1$-penalized EM estimation

In this section, we consider the implementation of the EM algorithm (Dempster et al., 1997) in the $\ell_1$-penalized likelihood approach. Green (1990) studied the convergence properties of the EM algorithm for the penalized likelihood. The EM algorithm is a popular method for maximum likelihood estimation in the case of incomplete data, which naturally occur in our setting as a result of the latent nature of $Z$.

The extended rank likelihood of $Y$ where $F_1, \ldots, F_p$ are considered nuisance parameters from Equation 2.4 is

$$L_Y(\Theta) = \int P(z \in D \mid \Theta) dz.$$  \hspace{1cm} (3.1)

For large sample sizes, the precision matrix $\Theta$ can be estimated by maximizing the log-likelihood $l(\Theta)$ as a function of $\Theta$. Whereas for high-dimensional data, we add an $\ell_1$-norm penalty to encourage sparsity in the precision matrix and the identification of the underlying graph. The $\ell_1$-penalized log-likelihood takes the form

$$l_\lambda(\Theta) = \log L_Y(\Theta) - \lambda \|\Theta\|_1,$$  \hspace{1cm} (3.2)

where the scalar parameter $\lambda \geq 0$ controls the size of the penalty and $\|\Theta\|_1 = \sum \sum_{i \neq j} |\Theta_{ij}|$.

Because of the complexity of maximizing the marginal log-likelihood $l_Y(\Theta)$ in Equations 3.1 and 3.2, we employ a Monte Carlo EM (MCEM) algorithm. This optimization strategy alternates iteratively between the E-step, computing conditional expectation of the complete log-likelihood

$$Q(\Theta \mid \hat{\Theta}^{(m)}) = E \left[ \log L_{Z,Y}(\Theta) \mid Y, \hat{\Theta}^{(m-1)} \right],$$

where $\hat{\Theta}^{(m-1)}$ is an estimate of $\Theta$ from the previous step of the algorithm and the complete likelihood of $(Z, Y)$ is given by the likelihood of $Z$, $L(\Theta; z)$, that is,

$$L_{Z,Y}(\Theta) = \prod_{i=1}^n \phi_p(z_i \mid \Theta) I_{z_i \in D(y)},$$  \hspace{1cm} (3.3)

such that

$$Q(\Theta \mid \hat{\Theta}^{(m-1)}) = E \left[ \log L(Z; \Theta) \mid z \in D(y), \hat{\Theta}^{(m-1)} \right],$$  \hspace{1cm} (3.4)

and the M-step, maximizing $Q(\Theta \mid \hat{\Theta}^{(m-1)})$, with a sparsity penalty $\lambda \|\Theta\|_1$, over $\Theta$.

The E-step and M-step are described in detail as follows.
**E-step:** Using the complete likelihood in Equation 3.3 into Equation 3.4, we obtain an additive constant

\[
Q(\Theta \mid \hat{\Theta}^{(m-1)}) = E \left[ \sum_{i=1}^{n} \left( \frac{1}{2} \log \det(\Theta) - \frac{1}{2} \frac{Z_i^T \Theta Z_i}{n} \right) \mid z_i \in D(y_i), \hat{\Theta}^{(m-1)} \right]
\]

\[
= \frac{n}{2} \left\{ \log \det(\Theta) - \frac{1}{n} \sum_{i=1}^{n} \text{Tr} \left( \Theta E \left[ Z_i Z_i^T \mid z_i \in D(y_i), \hat{\Theta}^{(m-1)} \right] \right) \right\}
\]

\[
= \frac{n}{2} \left\{ \log \det(\Theta) - \text{Tr} \left( \Theta \hat{R} \right) \right\},
\]

(3.5)

where \(\text{Tr}\) stands for the trace of a matrix and

\[
\hat{R} = \frac{1}{N} \sum_{i=1}^{n} E \left[ Z_i Z_i^T \mid z_i \in D(y_i), \hat{\Theta}^{(m-1)} \right].
\]

Note that the E-step involves an analytically intractable expectation, which we estimate by means of the Monte Carlo method of Wei and Tanner (1990). Alternatively, implementation of a Robbins–Monro type stochastic approximation version of the EM algorithm could be possible (see, for example, Delong et al., 1999; Jank, 2006). In the MCEM approach, the E-step is divided into a simulation and a Monte Carlo integration step as follows.

**Simulation step:** generate \(N(m)\) realizations \(z^{(m)}(s), s = 1, \ldots, N(m)\) from the conditional distribution \(\phi_p(z \mid z \in D_p(y), \hat{\Theta}^{(m-1)})\).

**Monte Carlo integration step:** compute the current approximation of \(Q(\Theta \mid \hat{\Theta}^{(m-1)})\) by the Monte Carlo average

\[
\tilde{Q}(\Theta \mid \hat{\Theta}^{(m-1)}) = \frac{1}{N(m)} \sum_{s=1}^{N(m)} \log L(z^{(m)}(s); \hat{\Theta}^{(m-1)}).
\]

(3.6)

Using similar earlier derivation, we obtain that

\[
\tilde{Q}(\Theta \mid \hat{\Theta}^{(m-1)}) = \frac{n}{2} \left\{ \log \det(\Theta) - \text{Tr} \left( \Theta \hat{R} \right) \right\},
\]

(3.7)

where

\[
\hat{R} = \frac{1}{N(m)} \sum_{s=1}^{N(m)} \left[ \frac{1}{n} \sum_{i=1}^{n} \left( z^{(m)}_i(s) \right) \left( z^{(m)}_i(s) \right)^T \right].
\]

Although the Monte Carlo approximation provides a solution to overcome the intractable E-step, it also introduces a Monte Carlo error whose magnitude depends on the size of \(N(m)\). It is recommended to start with small values of \(N(m)\) and increase this as the algorithm moves closer to convergence. Later in this section, we discuss how to determine the increase in \(N(m)\).
**M-step:** this involves updating the parameter $\Theta$ by maximizing the penalized $Q$-function for some values of the tuning parameter $\lambda$:

$$Q_p(\Theta \mid \hat{\Theta}^{(m-1)}) = Q(\Theta \mid \hat{\Theta}^{(m-1)}) - \lambda \|\Theta\|_1$$  \hspace{1cm} (3.8)

or the Monte Carlo approximation given by

$$\tilde{Q}_p(\Theta \mid \hat{\Theta}^{(m-1)}) = \tilde{Q}(\Theta \mid \hat{\Theta}^{(m-1)}) - \lambda \|\Theta\|_1.$$ \hspace{1cm} (3.9)

Using Equations 3.9 and 3.7, we obtained the current value as

$$\hat{\Theta}^{(m)} = \arg\max_\Theta \{\log\det(\Theta) - \text{Tr}(\Theta \tilde{R}) - \lambda \|\Theta\|_1\}.$$ \hspace{1cm} (3.10)

The maximization problem in Equation 3.10 takes the form of $c_\ell$-penalized likelihood for the Gaussian graphical models, and computation is carried out efficiently using the graphical lasso algorithm (Friedman *et al.*, 2008). This algorithm is fast and allows re-using the estimate under the previous value of the tuning parameter as a ‘warm’ start for the current value. The determination of the optimal value of $\lambda$ is discussed in the next section.

For a fixed value of the tuning parameter $\lambda$, to determine the increase in $N(m)$ and to control the convergence of the MCEM we consider the ascent-based MCEM, proposed by Caffo *et al.* (2005), that implements an automated data-driven strategy using the Monte Carlo resources across EM iterations. The basic idea of the penalized version of ascent-based MCEM is as follows.

Assume that $\{z^{(m)}(s); s = 1, \ldots, N(m)\}$ is a random sample from the conditional distribution $\phi_p(z \mid z \in D_p(y), \hat{\Theta}^{(m)})$. To control the ascent (increasing likelihood) property of the penalized likelihood with each MCEM iteration, we need to verify that

$$\Delta Q_p(\hat{\Theta}^{(m)}, \hat{\Theta}^{(m-1)}) = Q_p(\hat{\Theta}^{(m)} \mid \hat{\Theta}^{(m-1)}) - Q_p(\hat{\Theta}^{(m-1)} \mid \hat{\Theta}^{(m-1)}) > 0,$$ \hspace{1cm} (3.11)

where Equation 3.11 is computed based on a random sample $\{z^{(m)}(s); s = 1, \ldots, N(m)\}$ from $\phi_p(z \mid z \in D_p(y), \hat{\Theta}^{(m)})$.

Substituting Equation 3.4 into Equation 3.8, we obtain for Equation 3.11

$$\Delta Q_p(\hat{\Theta}^{(m)}, \hat{\Theta}^{(m-1)}) = E \left[ \log \left( \frac{L(Z; \hat{\Theta}^{(m)} \mid z \in D(y))}{L(Z; \hat{\Theta}^{(m-1)} \mid z \in D(y))} \right) \right]$$

$$- \lambda \left\{ \|\hat{\Theta}^{(m)}\|_1 - \|\hat{\Theta}^{(m-1)}\|_1 \right\},$$ \hspace{1cm} (3.12)

and we estimate $\Delta Q_p(\hat{\Theta}^{(m)}, \hat{\Theta}^{(m-1)})$ by the Monte Carlo approximation in Equations 3.6 and 3.9 as
\[
\Delta \tilde{Q}_p(\hat{\theta}^{(m)}, \hat{\theta}^{(m-1)}) = \tilde{Q}_p(\hat{\theta}^{(m)} | \hat{\theta}^{(m-1)}) - \tilde{Q}_p(\hat{\theta}^{(m-1)} | \hat{\theta}^{(m-1)}) \\
= \frac{1}{N(m)} \sum_{s=1}^{N(m)} \log \left( \frac{L(z^{(m)}(s); \hat{\theta}^{(m)})}{L(z^{(m)}(s); \hat{\theta}^{(m-1)})} \right) - \lambda \left\{ \| \hat{\theta}^{(m)} \|_1 - \| \hat{\theta}^{(m-1)} \|_1 \right\}
\]

Using the consistency of the \( \ell_1 \)-penalized estimator in Equation 3.10 (see, for example, Lam and Fan, 2009; Rothman et al., 2008), the central limit theorem combined with the continuous mapping theorem, or standard Taylor series type arguments, we obtained that the differences,

\[
\Delta \Delta Q_p = \Delta \tilde{Q}_p(\hat{\theta}^{(m)}, \hat{\theta}^{(m-1)}) - \Delta Q_p(\hat{\theta}^{(m)}, \hat{\theta}^{(m-1)}) \\
= \frac{1}{N(m)} \sum_{s=1}^{N(m)} \log \left( \frac{L(z^{(m)}(s); \hat{\theta}^{(m)})}{L(z^{(m)}(s); \hat{\theta}^{(m-1)})} \right) - E \left[ \log \left( \frac{L(Z; \hat{\theta}^{(m)} | z \in D(y))}{L(Z; \hat{\theta}^{(m-1)} | z \in D(y))} \right) \right],
\]

with a suitable normalization

\[
\sqrt{N(m)} \Delta \Delta Q_p = \frac{1}{N(m)} \sum_{s=1}^{N(m)} \log \left( \frac{L(z^{(m)}(s); \hat{\theta}^{(m)})}{L(z^{(m)}(s); \hat{\theta}^{(m-1)})} \right) - E \left[ \log \left( \frac{L(Z; \hat{\theta}^{(m)} | z \in D(y))}{L(Z; \hat{\theta}^{(m-1)} | z \in D(y))} \right) \right],
\]

have a limiting normal distribution with mean 0 and variance \( \frac{\sigma^2}{N(m)} \), where \( \sigma^2 \) is the variance of

\[
\Psi^{(m)}(s) = \log \left( \frac{L(z^{(m)}(s); \hat{\theta}^{(m)})}{L(z^{(m)}(s); \hat{\theta}^{(m-1)})} \right), \quad s = 1, \ldots, N(m).
\]

An asymptotic lower bound for the change, \( \Delta \tilde{Q}_p(\hat{\theta}^{(m)}, \hat{\theta}^{(m-1)}) \), between consecutive iterations is given by

\[
L_B = \Delta \tilde{Q}_p(\hat{\theta}^{(m)}, \hat{\theta}^{(m-1)}) - z_{\alpha} \frac{\sigma}{\sqrt{N(m)}}, \quad (3.13)
\]

such that \( L_B \) will be smaller than \( \Delta \tilde{Q}_p(\hat{\theta}^{(m)}, \hat{\theta}^{(m-1)}) \) with probability \( 1 - \alpha \) as \( N(m) \to \infty \).

If \( L_B > 0 \), there is sufficient evidence to conclude that \( \hat{\theta}^{(m)} \) increases the likelihood, \( \hat{\theta}^{(m)} \) is accepted as the \( m \)th parameter update, and the algorithm moves to the next iteration \( m + 1 \). If \( L_B < 0 \), the estimate \( \hat{\theta}^{(m)} \) is rejected, and we generate a larger Monte Carlo sample to estimate \( Q_p \) accurately. In this case, the \( m \)th iteration is repeated until \( L_B > 0 \). Following Caffo et al. (2005), we consider a geometric rate of increase in the Monte Carlo sample size; that is, when \( L_B < 0 \), let the Monte Carlo sample size be \( N(m) + N(m)/r \), for some \( r = 2, 3, \ldots \), such that \( N(m)/r \) Monte Carlo samples are drawn and appended to the current sample. We refer to Caffo et al. (2005) for an alternative

© 2015 The Author. Statistica Neerlandica © 2015 VVS.
criterion on how to specify the increase in the Monte Carlo sample size based on Type I and Type II error rates.

Similarly, we use the change in the penalized $Q$-function to assess convergence of the MCEM algorithm. Because in our high-dimensional setting the parameter set is large, stopping the MCEM algorithm when the marginal likelihood stabilizes or the change in the penalized $Q$-function is too small to detect is a convenient stopping rule. Define the stopping criterion by

$$S_B = \Delta \tilde{Q}_p(\hat{\Theta}^{(m)}, \hat{\Theta}^{(m-1)}) + z_\gamma \frac{\hat{\sigma}}{\sqrt{N(m)}},$$

such that $S_B$ is larger than $\Delta \tilde{Q}_p(\Theta^{(m)}, \Theta^{(m-1)})$ with probability $1 - \gamma$ as $N(m) \to \infty$. Thus, given a small specified constant $c$, we argue that convergence of the MCEM algorithm is reached when $S_B < c$. A small simulation study is included in the Appendix (Example 1).

### 3.2 Model selection

For high-dimensional data, the empirical covariance matrix is singular and poses computational problems. However, the $\ell_1$-penalized approach guarantees with probability one a positive definite precision matrix with the additional sparse property for large $\lambda$ values.

Under the $\ell_1$-penalized maximum likelihood setting, the sparsity of the estimated precision matrix is controlled by the penalty parameter $\lambda$ in Equation 3.10. We follow Yuan and Lin (2007) in using the Bayesian information criterion (BIC) in order to obtain a reasonably sparse precision matrix. One could also use other information criteria-based methods or cross-validation for the choice of $\lambda$. We will not consider those methods in this article. The BIC can be calculated as

$$BIC(\lambda) = -2l_Y(\hat{\Theta}_\lambda) + \frac{s(\hat{\Theta}_\lambda)}{2} \log n \approx C - n \left\{-\log \det(\hat{\Theta}_\lambda) + \text{Tr}(\hat{\Theta}_\lambda \hat{R}_{\hat{\Theta}_\lambda})\right\} + \frac{s(\hat{\Theta}_\lambda)}{2} \log n,$$

where $C$ is some constant, $s(\hat{\Theta}_\lambda)$ is the number of non-zero off-diagonal elements of $\hat{\Theta}_\lambda$, and $\hat{\Theta}_\lambda$ and $\hat{R}_{\hat{\Theta}_\lambda}$ are the final outputs of the iterative MCEM algorithm. The approximation made in the BIC is the result of a Laplace-type of approximation, which makes fast calculation feasible. We choose the optimal value of the penalty parameter that minimizes $BIC(\lambda)$ on a grid of candidate values for $\lambda$.

### 4 Data analysis

In this section, we return to the motivation of our methodological development and apply the proposed copula-based $\ell_1$-penalized EM estimation to the American National Football League and breast cancer data, which we introduced in Section 1.
Table 1. Variables observed during the 2011 regular NFL season

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>tyds</td>
<td>Total yards</td>
</tr>
<tr>
<td></td>
<td>1std</td>
<td>First downs</td>
</tr>
<tr>
<td></td>
<td>3rdmd</td>
<td>Third-down conversions</td>
</tr>
<tr>
<td></td>
<td>4thatt</td>
<td>Fourth-down attempts</td>
</tr>
<tr>
<td></td>
<td>4thpct</td>
<td>Fourth-down conversion %</td>
</tr>
<tr>
<td></td>
<td>rzatt</td>
<td>Red zone attempts</td>
</tr>
<tr>
<td></td>
<td>rzpct</td>
<td>Red zone conversion %</td>
</tr>
<tr>
<td></td>
<td>penyds</td>
<td>Penalty yards</td>
</tr>
<tr>
<td>Passing</td>
<td>pycomp</td>
<td>Passing completion</td>
</tr>
<tr>
<td></td>
<td>ppct</td>
<td>Passing completion %</td>
</tr>
<tr>
<td></td>
<td>pypatt</td>
<td>Passing yards per attempts</td>
</tr>
<tr>
<td></td>
<td>ptd</td>
<td>Passing touchdowns</td>
</tr>
<tr>
<td></td>
<td>PQBr</td>
<td>Passer rating</td>
</tr>
<tr>
<td>Rushing</td>
<td>ratt</td>
<td>Rushing attempts</td>
</tr>
<tr>
<td></td>
<td>rysds</td>
<td>Rushing yards</td>
</tr>
<tr>
<td></td>
<td>r.lt</td>
<td>Rushing touchdowns</td>
</tr>
<tr>
<td>Specials</td>
<td>psacks</td>
<td>Sacks</td>
</tr>
<tr>
<td></td>
<td>punts</td>
<td>Punts</td>
</tr>
<tr>
<td></td>
<td>avypp</td>
<td>Average yards per punt</td>
</tr>
<tr>
<td></td>
<td>puntblk</td>
<td>Punts blocked</td>
</tr>
<tr>
<td></td>
<td>puntRet</td>
<td>Punt return yards</td>
</tr>
<tr>
<td></td>
<td>puntTB</td>
<td>Punts for touchbacks</td>
</tr>
</tbody>
</table>

Notes: Each variable is measured from offensive (O) perspective and defensive (D) perspective during each match. The aim is to find the underlying conditional dependence structure between these categorical, count, and continuous variables.

4.1 Analysis of the National Football League data

We consider data from the 2011 regular season of the NFL available at the NFL website, which consists of $p = 74$ variables tagged as offense and defense from 32 teams. To account for the offense-defense perspective, a main strategy followed in organizing a team in the NFL; team performance variables listed in the following apply to offensive team (prefixed by O) and defensive team (prefixed by D). The only exceptions are percentage of wins (winper), points scored (pts), offense time of possession (Otop), and special team variables. The variables are listed in Table 1. Cohea and Payton (2011) analyzed NFL data from 2004 to 2008 using principal component and logistic regressions. They identified 14 variables, of which the most important predictors of game outcome were turnovers, rushing performance, passing, and total offense/defense.

We give a brief description of the NFL game. A single game in the NFL is played by two teams with 11 players each. The team on offense is given possession of the ball and attempts to take the opportunity to advance closer to the score line or to actually score in a series of plays called downs. Four downs are available to the offense...
team to cross at least 10 yards of the field. If this is not achieved, the ball goes to the opposing side. If the offensive team believes that it will not reach the 10th yard after the third down, it may use its fourth down to punt the ball by kicking it as far into the opposing team’s side as possible. There are two possibilities to score: a ‘touchdown’ stands for a goal where one member of the offensive team with possession of the ball crosses the goal line. This results in six points. A ‘field goal’ is a ball kicked between the goalposts of the opposing team and results in three points. On the other hand, the team on defense tries to prevent the offense from advancing or scoring by means of a ‘sack’ or an ‘interception’. A sack is a term used for a tackle by the defensive team, which brings down the offensive team player in possession of the ball before passing the ball to another offensive player. An interception is when the offensive team loses the control of the ball or ‘fumbles’ and results in an unintentional return of the ball to the defensive team. If during a down, the defensive team intercepts the ball from the offensive team, they can take the opportunity to score. A play stops when a player with the ball is brought down or when a scoring attempt has been made.

In a single NFL game, a team that scored the highest points wins the game. The winning points differ from game to game. Winning teams that scored points much higher than their opponents may be seen as having relatively strong offensive compared with their opponent team. Similarly, teams that do not allow more points to be scored may be considered as having a strong defense. Therefore, in the analysis that follows, we consider points scored by a team as a second possible game outcome variable in addition to the winning percentage. The list of variables in Table 1 includes mixed variables that are a collection of binary, count, and continuous variables.

We apply the proposed $\ell_1$-penalized ascent-MCEM algorithm using the Gaussian copula graphical models on this dataset to reconstruct the network displaying interdependence among the set of variables in Table 1. The BIC results in an optimal $\lambda = 0.12$. Mainly, we are interested on the subset of variables directly related to the percentage of winning and points scored, which is extracted from a complex dependence pattern among the observed variables induced by the underlying multivariate normal latent variables and is shown in Figure 1.

In Figure 1, we systematically presented three groups of variables and their relationship within and across the groups. The first group of variables displayed in the inner circle of Figure 1 shows the expected positive relationship between the two game outcome variables: winning percentage and points scored. These outcome variables are directly related to game performance variables included in the middle circle. Further, the game performance variables inside the middle circle are relate to each other and directly to game performance variables in the outer circle. That is also to say that the game outcome variables in the inner circle are indirectly related to those variables in the outer circle of Figure 1.

Further exploration of the links in the network helps us to understand the interplay of the game performance and game outcome variables. Here, to demonstrate the interdependence relationship displayed by the network in Figure 1, we give an illustration by highlighting the links between the outcome variables: winper and pts and
Fig. 1. Conditional dependence subnetwork of team game performance and game outcome variables for the National Football League 2011 season data with partial correlations higher than |0.05|. Bold solid lines represent negative conditional associations, and light solid lines represent positive conditional associations.

Passer rating (OPQBr) and fourth-down attempts (O4thatt) from the offensive perspective and turnovers (Dpint) and rushing (that includes Dratt and Drtd) from the defensive perspective.

Passer rating is a measure of performance of quarterbacks or any other passers in the NFL and is calculated based on passing completion percentage, passing yards, touchdowns, and interceptions. Similarly, in the reconstructed network in Figure 1, passer rating is related to passing yards per attempt (Opypatt), passing and rushing touchdowns (Optd and Ortd), fourth-down conversion (O4thPct), and interception (Opint), but not directly to passing completion percentage. Passer rating is also positively related to both game outcome variables and O4thPct. This suggests that passing efficiency of a team on the offensive side is crucial to score more points and win a game. On the other hand, fourth-down attempt or the use of last chance in the series of plays by the offensive team is negatively associated with game outcomes and a sign of weakness for a team on the offensive side.

From a defensive perspective, on one hand, turnovers (Dpint) or takeaways by the defensive team are positively related to game outcomes that indicate the strength of the team on the defense side, and on the other hand, defense rushing (Dratt and Drtd) has a negative relationship with game outcomes. It shows that a defense team that is unable to stop the offensive players from advancing and scoring is a sign of weakness of the team. It is also found that defense rushing is negatively related to red zone attempt (Orzatt)
and time of possession (Otop) of the offensive team. In general, a careful understanding of the links in the network helps to generate interesting strategies that can be useful for managers, players, and other stakeholders involved in American football.

4.2 Analysis of breast cancer data

We next apply the proposed approach to the breast cancer data, which we introduced in section 1. The breast cancer experiment is a clinical study of DNA amplification and deletion patterns, using microarray technology. Its aim is to study the relationship between DNA amplification and deletion patterns and the severity of the breast cancer, as measured by several clinical indicators on the patients. The data from the breast cancer experiment include 59 selected genes and six clinical variables obtained from 62 breast cancer patients. A brief description of clinical and genomic variables included in this study is presented in Table 2.

In the breast cancer study, missing data rates among each of the gene amplification variables were less than 3%. Some parts of the array may have been damaged and resulted in some data excluded from consideration. Similarly, the missing data rates for the clinical variables, for example, breast cancer tumor size and grade, were 8% and 19%, respectively. We consider data missing in this study as missing completely at random so that the missing values are easily determined from the latent variable distribution defined on the interval \((−\infty, \infty)\) in the E-step of the EM algorithm.

We apply the proposed approach to the breast cancer data to investigate the dependence pattern among breast cancer survival, genetic and clinical variables. The BIC resulted in an optimal penalty value of \(\lambda = 0.2\). A subnetwork of the complex dependence pattern among the observed variables induced by the underlying multivariate normal latent variables is displayed in Figure 2. This subnetwork mainly includes clinical and genomic variables directly related to breast cancer survival. In Figure 2, positive relationships are represented by bold solid lines and negative relationships by light solid lines.

As can be seen in Figure 2, NPI is directly related to breast cancer tumor size, cancer grade, and number of nodes involved, that is, the higher the NPI score, the more aggressive the breast cancer and higher the chance of death due to breast cancer. Moreover, breast cancer survival is directly related to gene BCL2 \(^3\)'(B-cell lymphoma 2). That is, deletion of BCL2 \(^3\)' gene is related to high risk of breast cancer death, which

<table>
<thead>
<tr>
<th>Table 2. List of genomic and clinical variables obtained from the breast cancer experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at diagnosis</td>
</tr>
<tr>
<td>Size</td>
</tr>
<tr>
<td>Survival status</td>
</tr>
<tr>
<td>Grade</td>
</tr>
<tr>
<td>Nodal stage</td>
</tr>
<tr>
<td>NPI</td>
</tr>
<tr>
<td>Genes</td>
</tr>
</tbody>
</table>

Notes: The aim is to find the underlying conditional dependence structure between these categorical, count, and continuous variables.

NPI, Nottingham prognostic index.
is consistent with the literature that BCL2 family of genes are well-known inhibitor of cancer cell growth. For instance, BCL2 has been demonstrated as a single marker in breast cancer in Dawson et al. (2010), and BCL2 expression has been invariably found to correlate with a better prognosis of breast cancer in Martinez-Arribas et al. (2007). Moreover, BCL2 proteins are highly promising targets for improving and developing novel personalized cancer treatment strategies (Vogler, 2014; Gibson et al., 1999). On the other hand, in Figure 2, BCL2 3′ is negatively associated with the amplification of gene ESR (estrogen receptor), which is further directly related to the size of breast cancer tumor (SIZE). The relationship between the genes BCL2 3′ and ESR has also been investigated by (Lee et al., 2013). They suggest that targeting the interaction between BCL and ESR may enhance the efficacy of chemotherapeutic agents for patients with advanced breast cancer treatment strategies (Lee et al., 2013).

Further, we see that the NPI score and the three clinical variables related to some selected genes. For example, NPI score is positively related to cyclin D1 (CCND1) and MYC genes; size of breast cancer is positively related to ESR, LAMC2 (laminin gamma 2), and PAK1 (p21 protein-activated kinase 1) genes, and tumor grade is positively related to ERBB2 (erb-b2 receptor tyrosine kinase 2). The clinical relevance of these genes has been demonstrated by various clinical studies and many different targeted therapies like hormone therapies, gene expression modulator and apoptosis inducer have been approved for use in breast cancer treatment, such as BCL2 (Vogler, 2014), CCND1 (Xu et al., 2013), MYC (Singhi et al., 2012), ESR (Holst et al., 2012),
Despite the rapid growth in technology and the availability of various breast cancer treatment strategies, cancer treatment still faces several challenges. The findings in this study show the interactions between the clinical and genomic variables in breast cancer. These interactions suggest that combining clinical variables with genomic information can lead to more refined treatment approaches at the individual patient level. Thus, further investigation towards a complete picture of interactions between clinical and genomic variables is a promising approach to address the challenges in breast cancer treatment strategies.

5 Conclusions

Large high-dimensional datasets have become a common feature of many modern measurement techniques. Many of these datasets are generated without any specific research question in mind, and exploration of the multivariate structure of the data can be used for generating new hypotheses. In this article, we have seen how we can use \( \ell_1 \)-penalized graphical models to study conditional independence for a large set variable of arbitrary type. We also demonstrated that entries of the precision matrix of the Gaussian copula are closely related to conditional local log-odds ratios of ordered discrete variables. The accuracy of the approximated relation improves with increasing the number of categories of ordered discrete variables. In practice, this could happen when a set of mixed variables under investigation involves continuous, count, and ordered categorical variables with many levels. In particular, the relationship suggests that zero entries in the precision matrix of a Gaussian copula imply near conditional independence for mixed variables. Our motivating examples were datasets on game performance measures during the 2011 regular NFL season and genomic and clinical measurements on breast cancer. We obtained a sparse representation of the conditional dependencies between the variables under considerations, which generated several interesting hypotheses on the importance of the selected variables for winning a football game and for more accurate diagnostics and treatment of breast cancer at individual patient level.

Acknowledgements

The authors are grateful to the Editor, an Associate Editor, and anonymous referees for their valuable comments that considerably improved the manuscript.

Appendix

Example: \( \ell_1 \)-penalized ascent-MCEM

This example demonstrates the convergence of the \( \ell_1 \)-penalized ascent-MCEM. Consider small sample \((n = 10)\) simulated data from a multivariate normal distribution.
with $p = 10$ variables having a random graph structure with a sparse precision matrix generated using the R package \texttt{flare}. Five of the variables are discretized to form ordered variables, each having five levels. We apply the proposed $\ell_1$-penalized ascent-MCEM approach on the 10 mixed continuous and ordinal variables for the tuning parameter $\lambda = 0.4$. The initial Monte Carlo sample size is set at 100, but during, for example, the first 15 iterations, the Monte Carlo sample sizes increase between 300 and 500 to keep the increase in the marginal penalized log-likelihood with probability 0.95. The average values for the penalized $Q$-functions or marginal log-likelihood paths obtained from 20 simulations are displayed in Figure A1. We see that for the ascent-MCEM, the paths in Figure A1(a) increase and stabilize within a few iterations, suggesting the convergence of the algorithm. On the other hand, an MCEM with a fixed Monte Carlo sample of size of 100 results in more unstable paths of the penalized marginal log-likelihood compared with the ascent-MCEM within the 15 iterations; see Figure A1(b).

**Supporting information**

Additional supporting information may be found in the online version of this article.  

**Supplement R code** (http://www.math.rug.nl/stat/Main/Software). The computer code used in this article is available online as an R statistical computing environment package, \texttt{gcgmmd}.

**References**


© 2015 The Author. Statistica Neerlandica © 2015 VVS.


Received: March 2013. Revised: March 2015.