

University of Groningen

Extensions of graphical models with applications in genetics and genomics

Behrouzi, Pariya

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version

Publisher's PDF, also known as Version of record

Publication date:
2018

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Behrouzi, P. (2018). *Extensions of graphical models with applications in genetics and genomics*. [Thesis fully internal (DIV), University of Groningen]. University of Groningen.

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Chapter 5

Dynamic Chain Graph Models for Ordinal Time Series Data

Abstract

This chapter introduces sparse dynamic chain graph models for network inference in high dimensional non-Gaussian time series data. The proposed method is parametrized by a precision matrix that encodes intra time-slice conditional independences among variables at a fixed time point, and an autoregressive coefficient that contains dynamic conditional independence interactions among time series components across consecutive time steps. The proposed model is a Gaussian copula vector autoregressive model, which is used to model sparse interactions in a high-dimensional setting. Estimation is achieved via a penalized EM algorithm. In this chapter we use an efficient coordinate descent algorithm to optimize the penalized log-likelihood with the smoothly clipped absolute deviation penalty. We demonstrate our approach on simulated and psychological datasets. Our method is implemented in an R package **tsnetwork**.

Key words: Chain graph models; time-series data; Latent variable; Gaussian copula; SCAD penalty ; L_1 penalty; penalized likelihood; Vector autoregressive model.

5.1 Introduction

Graphical models are an efficient tool for modeling and inference in high dimensional settings. Directed acyclic graph (DAG) models, known as Bayesian networks (Lauritzen, 1996), are often used to model asymmetric cause-effect relationships. Models represented

by undirected graphs are used to model symmetric relationships, like gene regulatory networks.

Some graphical models can represent both asymmetric and symmetric relationships simultaneously. One such model is the so-called chain graph model (Lauritzen, 1996; Lauritzen and Wermuth, 1989) which is a generalization of directed and undirected graphical models. Chain graph models contain a mixed set of directed and undirected edges. The vertex set of a chain graph can be partitioned into chain components in which edges within a chain component are undirected, whereas the edges between two chain components are directed and point in the same direction. Recently, chain graph models have been considered in a time series setting (Abegaz and Wit, 2013; Gao and Tian, 2010; Dahlhaus and Eichler, 2003). There is rich literature on the reconstruction of undirected graphs for continuous data, categorical data, and mixed categorical-and-continuous data (Behrouzi and Wit, 2017; Mohammadi et al., 2015; Dobra et al., 2011; Hoff, 2007) as well as for directed acyclic graphs (Colombo et al., 2012; Kalisch and Bühlmann, 2007). Recently, Abegaz and Wit (2013) proposed a method based on a chain graph model for analyzing time-course continuous data, like gene expression data. However, many real-world time series data are not continuous, but are categorical or mixed categorical-and-continuous. Until now the construction of dynamic networks for non-continuous time series data has remained unexplored. Here, we develop a method to explore dynamic or delayed interactions and contemporaneous interactions for time series of categorical data and time series of mixed categorical-and-continuous data.

The proposed method is based on chain graph models, where the ordered time steps build a DAG of blocks, each of which contains an undirected network of the variables under consideration at that time point. The method developed in this chapter is designed to analyze the nature of interactions present in repeated multivariate time-series of mixed categorical-and-continuous data, where we use time series chain graphical models to study the conditional independence relationships among variables at a fixed time point as well as “causal” relationships among time series components across consecutive time steps. We use the concept of Granger causality (Granger, 1969), which exploits the natural time ordering to achieve a “causal” ordering of the variables in multivariate time series. The idea of this causality concept is based on predictability, where one time series is said to be Granger causal for another series if the latter series can be better predicted using all available information, rather than only the information from the latter series had been used. Our inference procedure not only enforces sparsity on interactions within each time step, but also between time steps; this feature is particularly realistic in a real-world dynamic

networks setting.

We proceed as follows: in section 5.2 we explain the method: we first introduce dynamic chain graph models in section 5.2.1, and then propose the Gaussian copula for mixed scale time series data in section 5.2.2. In sections 5.2.3 and 5.2.4 we define a model for underlying multivariate time series components, and we explain the procedure of penalized inference based on the L1 norm and smoothly clipped absolute deviation (SCAD) penalty terms. In section 5.2.5 we present a method for obtaining the log-likelihood of the observed mixed scale time series component under the penalized EM algorithm, and we proceed with model selection for tuning the penalty terms. In section 5.3 we study the performance of the proposed dynamic chain graph model under different scenarios. We also compare its performance with the other available methods. In section 5.4 we demonstrate the use of the proposed method in investigating the course of depression and anxiety disorders.

5.2 Methods

5.2.1 Dynamic chain graph models

A chain graph is defined as $G = (V, E)$ where V is a set of vertices (nodes) and E is a set of ordered and unordered pairs of nodes, called edges, which contains the directed and undirected interactions between pairs of nodes. A dynamic chain graph model is associated with a time series chain graph model, where the dependence structure of the time series components can be divided into two sets: *intra time-slice dependencies*, represented by undirected edges that specify the association among variables in a fixed time step, and *inter time-slice dependencies*, represented by associations among variables across consecutive time steps. Links across time steps are directed, pointing from a set of nodes at a previous time step, $V_{(t-1)}$, to nodes at the current time step, V_t . The dynamic chain graph model in our modeling framework relates the time series components at time t only to that of time $t - 1$, but this can be easily extended to a higher order ($d \geq 2$) of time steps.

Let $\mathbf{Y}(t) = (Y_1(t), \dots, Y_p(t))'$, $t = 1, \dots, T$ be an p -dimensional time series vector representation of p variables that have been studied longitudinally across T time points. Each time series component $Y(t)$ is assumed to be sampled n times. Thus, $Y_{ij}(t)$ represents the value of the j -th variable at time t for the i -th sample, $i = 1, \dots, n$, $j = 1, \dots, p$.

Here, we focus on non-Gaussian multivariate time series data such as ordinal-valued time series, taking values in $\{0, 1, \dots, (c_k - 1)\}$, where c_k is the number of possible categories, or mixed categorical-and-continuous time series data, as routinely occurring in real world

settings.

5.2.2 Gaussian Copula

To model dependencies among p -dimensional vector y we use a Gaussian copula, defined as

$$F(y_1, \dots, y_p) = \Phi_p\left(\Phi^{-1}(F_1(y_1)), \dots, \Phi^{-1}(F_p(y_p)) \middle| \Omega_{p \times p}\right) \quad (5.1)$$

where $\Phi_p(\cdot | \Omega)$ is the a p -dimensional Gaussian cdf with correlation matrix $\Omega_{p \times p}$, and $y = (y_1, \dots, y_p)$. From equation (5.1) the following properties are clear: the joint marginal distribution of any subset of Y has a Gaussian copula with a correlation matrix Ω and univariate marginals F_j . The Gaussian copula can be expressed in terms of a latent Gaussian variable $\mathbf{Z} = Z_1, \dots, Z_p$ as follows

$$\mathbf{Z} \sim \mathcal{N}(0, \Omega_{p \times p})$$

and

$$Y_j = F_j^{-1}(\Phi(Z_j)). \quad (5.2)$$

Since the marginal distributions F_j are nondecreasing, observing $y_{i_1j} < y_{i_2j}$ implies $z_{i_1j} < z_{i_2j}$. This can be written as set $\mathcal{A}(y)$ where, given the observed data $y_j = (y_{1,j}, \dots, y_{n,j})$, the latent samples $z_j = (z_{1,j}, \dots, z_{n,j})$, are constrained to belong to the set

$$\mathcal{A}(y) = \{z \in R^{n \times p} : \max\{z_{s,j} : y_{s,j} < y_{r,j}\} < z_{r,j} < \min\{z_{s,j} : y_{r,j} < y_{s,j}\}\}$$

If an observed value of y_j is missing, we define the lower bound and the upper bound of $z_j^{(r)}$ as $-\infty$ and ∞ , respectively.

5.2.3 Model definition

We assume a stable dynamic chain graph model, meaning that the structure of interactions within each time point remains stable for previous and current time steps, and interactions between consecutive time steps are also stable. We use a vector autoregressive process of order 1, VAR(1),

$$Z_t = \Gamma Z_{(t-1)} + \epsilon_t \quad (5.3)$$

to describe the directed latent interactions, where $\epsilon_t \sim N(0, \Theta^{-1})$ describes the undirected instantaneous interactions.

The parameter set of this model contains all the conditional independence relationships in the dynamic chain graph model where the following terms hold: $\theta_{jj} = 0$ if and only if $Z_j^{(t)} \perp\!\!\!\perp Z_j^{(t)} \mid Z_{-j,j}^{(t)} Z^{(t-1)}$, and $\gamma_{jj} = 0$ if and only if $Z_j^{(t)} \perp\!\!\!\perp Z_j^{(t-1)} \mid Z_{-j}^{(t)} Z_{-j}^{(t-1)}$.

Given the set $\mathcal{A}(y)$, we calculate the likelihood as

$$\begin{aligned} f(\mathbf{y} \mid \Theta, \Gamma, F) &= f(\mathbf{y}, \mathbf{z} \in \mathcal{A}(\mathbf{y}) \mid \Theta, \Gamma, F) \\ &= f_Z(\mathbf{z} \in \mathcal{A}(\mathbf{y}) \mid \Theta, \Gamma) f(\mathbf{y} \mid \mathbf{z} \in \mathcal{A}(\mathbf{y}), \Theta, \Gamma, F) \end{aligned} \quad (5.4)$$

where $y = \{(y_1^{(t)}, \dots, y_p^{(t)})\}_{t=1}^T$ and $F = \{(F_1^{(t)}, \dots, F_p^{(t)})\}_{t=1}^T$. Given the set of parameters, the event $\mathbf{z} \in \mathcal{A}(\mathbf{y})$ in (5.4) does not depend on marginals and contains the relevant information about the copula and the parameters of interest Θ and Γ . We drop the second term in (5.4) because it provides no information about intra and inter time-slice dependencies. As Hoff (2007) proposes, we use $f_Z(\mathbf{z} \in \mathcal{A}(\mathbf{y}) \mid \Theta, \Gamma)$ as the rank likelihood,

$$\begin{aligned} \ell_Y(\Theta, \Gamma) &= \sum_{i=1}^n \log f(\mathbf{z}_i \in \mathcal{A}(\mathbf{y}) \mid \Theta, \Gamma) \\ &= \sum_{i=1}^n \sum_{t=2}^T \log f(z_i^{(t)} \in \mathcal{A}(\mathbf{y}_i^{(t)}) \mid z_i^{(t-1)} \in \mathcal{A}(\mathbf{y}_i^{(t-1)}); \Theta, \Gamma) + \log f(z_i^{(1)} \in \mathcal{A}(\mathbf{y}_i^{(1)}) \mid \Theta, \Gamma) \end{aligned} \quad (5.5)$$

We ignore the second term in (5.5) as we do not want to make additional assumptions on the unconditional distribution of $Y^{(1)}$. We start from $t = 2$, where we compute the conditional log-likelihood using the conditional distribution $f(z^{(t)} \mid z^{(t-1)})$. According to (5.3) the conditional distribution $Z^{(t)} \mid Z^{(t-1)}$ follows a multivariate normal distribution

$$Z^{(t)} \mid Z^{(t-1)} = z^{(t-1)} \sim \mathcal{N}(\Gamma z^{(t-1)}, \Theta^{-1}) \quad (5.6)$$

Whose density for t -th observation is defined as

$$f(z^{(t)} \mid z^{(t-1)}; \Theta, \Gamma) = (2\pi)^{p/2} \det(\Theta)^{1/2} \exp \left[\frac{1}{2} \left(z^{(t)} - \Gamma z^{(t-1)} \right)' \Theta \left(z^{(t)} - \Gamma z^{(t-1)} \right) \right]. \quad (5.7)$$

5.2.4 Penalized EM inference

In Gaussian copula, we treat marginal distributions as nuisance parameters since our main goal is to learn the dependence structure among time series components both at a fixed time step $t \in \mathbb{N}$ and also across consecutive time steps. We use an empirical marginal cdf $\hat{F}_j = \frac{n}{n+1} \sum_{i=1}^n \frac{1}{n} 1(y_{ij} \leq y)$ (Genest et al., 1995) to estimate marginals.

Genetic time series data are often high dimensional due to a large number of variables that are measured on small numbers of samples across only a few time steps. Furthermore, many real-world networks (e.g. genetic, genomics, and brain networks) are intrinsically sparse. Thus, incorporating sparsity into the proposed dynamic chain graph model makes the derived model more biologically plausible. Accordingly, we propose a dynamic chain graph model for genetic data based on penalized likelihood. In order to find the penalized maximum likelihood estimation we will use the EM algorithm (Green, 1990). This modeling technique provides sparse estimates of the autoregressive coefficient matrix Γ and the precision matrix Θ in (5.3), which are used to reconstruct inter and intra time-slice conditional independences, respectively.

The E-step of the EM algorithm is given by

$$\begin{aligned} Q(\Theta, \Gamma \mid \Theta^*, \Gamma^*) &= E_z \left[\ell_{Y,Z}(\Theta, \Gamma) \mid y_i, \Theta^*, \Gamma^* \right] \\ &= E_z \left[\sum_{i=1}^n \sum_{t=2}^T \log f(Z_i^{(t)} \mid Z_i^{(t-1)}; \Theta, \Gamma) \mid y_i, \Theta^*, \Gamma^* \right]. \end{aligned} \quad (5.8)$$

Under the assumption described in (5.6), the E-step can be written as

$$Q(\Theta, \Gamma \mid \Theta^*, \Gamma^*) = \frac{n(T-1)}{2} \left[-p \log(2\pi) + \log \det(\Theta) - \text{tr} \left(E(S_\Gamma \mid y_i, \Theta^*, \Gamma^*) \Theta \right) \right] \quad (5.9)$$

where

$$\begin{aligned} E(S_\Gamma \mid y_i, \Theta^*, \Gamma^*) &= \frac{1}{n(T-1)} \sum_{i=1}^n \sum_{t=2}^T E_Z \left[(Z_i^{(t)} - \Gamma Z_i^{(t-1)})(Z_i^{(t)} - \Gamma Z_i^{(t-1)})' \mid y_i, \Theta^*, \Gamma^* \right] \\ &= \frac{1}{n(T-1)} \left[S_{cc} - S_{cp} \Gamma' - \Gamma S_{cp}' + \Gamma S_{pp} \Gamma' \right] \end{aligned} \quad (5.10)$$

such that conditional expectation at current time, S_{cc} , and at past time, S_{pp} , is defined as

$$S_{cc} = \sum_{i=1}^n \sum_{t=2}^T E_Z [Z_i^{(t)} Z_i^{(t)'} \mid y_i; \Theta^*, \Gamma^*], \quad S_{pp} = \sum_{i=1}^n \sum_{t=1}^{T-1} E_Z [Z_i^{(t)} Z_i^{(t)'} \mid y_i; \Theta^*, \Gamma^*]$$

and the conditional expectation at inter time-slice dependence is

$$S_{pc} = \sum_{i=1}^n \sum_{t=2}^T E_Z[Z_i^{(t-1)} Z_i^{(t)'} | y_i; \Theta^*, \Gamma^*].$$

The latent variables $Z_i^{(t-1)} = \{Z_{i,1}^{(t-1)}, \dots, Z_{i,p}^{(t-1)}\}$ and $Z_i^{(t)} = \{Z_{i,1}^{(t)}, \dots, Z_{i,p}^{(t)}\}$ are used to calculate the conditional expectation of intra time-slice dependencies S_{pp} and S_{cc} , respectively. And $Z_i^{(pc)} = \{Z_{i,1}^{(t-1)}, \dots, Z_{i,p}^{(t-1)}, Z_{i,1}^{(t)}, \dots, Z_{i,p}^{(t)}\}$ is used to calculate S_{pc} . All three above-mentioned conditional expectations are a $p \times p$ matrix. When $j = j'$ they can be computed through the second moment $E(Z_{ij}^{(t)2} | y_i; \Theta^*, \Gamma^*)$. When $j \neq j'$ we use a mean field theory approach (Chandler, 1987) to approximate them as

$$E\left(Z_{i,j}^{(t)} Z_{i,j'}^{(t)} \mid y_i; \Theta^*, \Gamma^*\right) \approx E\left(Z_{i,j}^{(t)} \mid y_i; \Theta^*, \Gamma^*\right) E\left(Z_{i,j'}^{(t)} \mid y_i; \Theta^*, \Gamma^*\right) \quad (5.11)$$

for intra time-slice dependencies, and for inter time-slice dependencies as follows:

$$E\left(Z_{i,j}^{(t-1)} Z_{i,j}^{(t)} \mid y_i; \Theta^*, \Gamma^*\right) \approx E\left(Z_{i,j}^{(t-1)} \mid y_i; \Theta^*, \Gamma^*\right) E\left(Z_{i,j}^{(t)} \mid y_i; \Theta^*, \Gamma^*\right) \quad (5.12)$$

This approximation performs well when the interaction between $Z_{i,j}^{(t)}$ and $Z_{i,j'}^{(t)}$ given the rest of the variables, and the interaction between $Z_{i,j}^{(t-1)}$ and $Z_{i,j}^{(t)}$ given the rest of the variables, are close to being independent; this often holds in our proposed dynamic chain graph model, in which Θ and Γ are sparse.

When $j \neq j'$ the off-diagonal elements of S_{cc} , S_{pp} , and S_{pc} matrices can be computed through the first moment as

$$E\left(Z_{i,j}^{(t)} \mid y_i; \Theta^*, \Gamma^*\right) = E\left[E\left(Z_{i,j}^{(t)} \mid Z_i^{(t-1)}, Z_{i,-j}^{(t)}, Z_i^{(t+1)}, y_{i,j}^{(t)}; \Theta, \Gamma\right) \mid y_i; \Theta^*, \Gamma^*\right] \quad (5.13)$$

and the second moments as

$$E\left(Z_{i,j}^{(t)2} \mid y_i; \Theta^*, \Gamma^*\right) = E\left[E\left(Z_{i,j}^{(t)2} \mid Z_i^{(t-1)}, Z_{i,-j}^{(t)}, Z_i^{(t+1)}, y_{i,j}^{(t)}; \Theta, \Gamma\right) \mid y_i; \Theta^*, \Gamma^*\right] \quad (5.14)$$

Given the property of Gaussian distribution, $(Z_i^{(t)}, Z_i^{(t+1)}) \mid Z_i^{(t-1)}; \Theta, \Gamma$ follows a multivariate normal distribution with mean and variance-covariance matrix

$$\mu = \begin{bmatrix} \Gamma z_i^{(t-1)} \\ \Gamma^2 z_i^{(t-1)} \end{bmatrix} \quad V = \begin{bmatrix} \Theta^{-1} & \Theta^{-1} \Gamma \\ \Gamma \Theta^{-1} & \Gamma \Theta^{-1} \Gamma' + \Theta^{-1} \end{bmatrix}.$$

Therefore, the conditional distribution of $Z_{i,j}^{(t)} \mid Z_i^{(t-1)}, Z_{i,-j}^{(t)}, Z_i^{(t+1)}; \Theta, \Gamma$ inside the inner expectation in (5.13) and (5.14) follows a multivariate normal distribution with mean μ_{ij} and variance v_{ij} as follows:

$$\mu_{ij} = (\Gamma_i z_i^{(t-1)})_j + V_{j,-j} V_{-j,-j}^{-1} \left(\begin{bmatrix} z_{i,-j}^{(t)} \\ z_i^{(t+1)} \end{bmatrix} - \begin{bmatrix} \Gamma_i z_i^{(t-1)} \\ \Gamma_i^2 z_i^{(t-1)} \end{bmatrix} \right)$$

$$v_{ij} = V_{j,j} - V_{j,-j} V_{-j,-j}^{-1} V_{-j,j}.$$

Calculating the exact value of the first and second moments is computationally expensive. Moreover, we approximate the first and the second moments as follows

$$E\left(Z_{i,j}^{(t)} \mid y_i; \Theta^*, \Gamma^*\right) \approx E\left[E\left(Z_{i,j}^{(t)} \mid Z_i^{(t-1)}, Z_{i,-j}^{(t)}, y_{i,j}^{(t)}; \Theta, \Gamma\right) \mid y_i; \Theta^*, \Gamma^*, \right] \quad (5.15)$$

$$E\left(Z_{i,j}^{(t)2} \mid y_i; \Theta^*, \Gamma^*\right) \approx E\left[E\left(Z_{i,j}^{(t)2} \mid Z_i^{(t-1)}, Z_{i,-j}^{(t)}, y_{i,j}^{(t)}; \Theta, \Gamma\right) \mid y_i; \Theta^*, \Gamma^*\right] \quad (5.16)$$

The conditional distribution of $Z_i^{(t)} \mid Z_i^{(t-1)}; \Theta, \Gamma$ follows a multivariate normal distribution with mean $\Gamma_i z_i^{(t-1)}$ and variance-covariance matrix Θ^{-1} . Due to a property of Gaussian distribution, the conditional distribution of $Z_{i,j}^{(t)} \mid Z_i^{(t-1)}, Z_{i,-j}^{(t)}; \Theta, \Gamma$; inside the inner expectation in (5.15) and (5.16) follows a multivariate normal distribution with mean and variance-covariance matrix as follows

$$\mu'_{i,j} = (\Gamma_i z_i^{(t-1)})_j + \widehat{\Sigma}_{j,-j} \widehat{\Sigma}_{-j,-j}^{-1} \left(z_{i,-j}^{(t)\tau} - (\Gamma_i z_i^{(t-1)})_{-j} \right)$$

$$\sigma'_{i,j}{}^2 = \widehat{\Sigma}_{j,j} - \widehat{\Sigma}_{j,-j} \widehat{\Sigma}_{-j,-j}^{-1} \widehat{\Sigma}_{-j,j}.$$

We remark that conditioning $z_{i,j}^{(t)}$ on $z_i^{(t-1)}, z_{i,-j}^{(t)}$ and $y_{i,j}^{(t)}$ is equivalent to

$$z_{i,j}^{(t)} \mid z_i^{(t-1)}, z_{i,-j}^{(t)}, c_{j,y_{i,j}^{(t)}} \leq z_{i,j}^{(t)} \leq c_{j,y_{i,j}^{(t)}+1}.$$

Thus, this conditional distribution follows a truncated normal on the interval $[c_{j,y_{i,j}^{(t)}}, c_{j,y_{i,j}^{(t)}+1}]$, where the first and second moments can be obtained via lemma 5.1.

Lemma 5.1. (Johnson et al., 1995). Let $Z \sim \mathcal{N}(\mu_0, \sigma_0^2)$ such that $\delta_1 = (c_1 - \mu_0)/\sigma_0$ and $\delta_2 = (c_2 - \mu_0)/\sigma_0$ are true for any constants that $c_1 < c_2$. Then the first and second moments

of the truncated normal distribution on the interval (c_1, c_2) are defined as

$$E(Z|c_1 \leq Z \leq c_2) = \mu_0 + \frac{\phi(\delta_1) - \phi(\delta_2)}{\Phi(\delta_2) - \Phi(\delta_1)}\sigma_0$$

$$E(Z^2|c_1 \leq Z \leq c_2) = \mu_0^2 + \sigma_0^2 + 2\frac{\phi(\delta_1) - \phi(\delta_2)}{\Phi(\delta_2) - \Phi(\delta_1)}\mu_0\sigma_0 + \frac{\delta_1\phi(\delta_1) - \delta_2\phi(\delta_2)}{\Phi(\delta_2) - \Phi(\delta_1)}\sigma_0^2$$

where $\phi(\cdot)$ is the density function of the standard normal distribution.

Both means $\mu_{i,j}$ and $\mu'_{i,j}$ are a linear function of $z_{i,-j}^{(t)}$, and both $\frac{\phi(\delta_1) - \phi(\delta_2)}{\Phi(\delta_2) - \Phi(\delta_1)}$ and $\frac{\delta_1\phi(\delta_1) - \delta_2\phi(\delta_2)}{\Phi(\delta_2) - \Phi(\delta_1)}$ are nonlinear functions of $z_{i,-j}^{(t)}$. Applying Lemma 5.1 to the conditional expectations in (5.15) and (5.16) leads to the approximations (5.24) and (5.25) in the Appendix.

Moreover, we approximate the elements of inter time-slice conditional expectation matrix S_{pc} through equations (5.24) and (5.25). For approximating the elements of intra time-slice conditional expectation matrices S_{pp} , S_{cc} we refer to the Appendix.

The M-step of the EM algorithm contains a two-stage optimization process where we maximize expectation of the penalized log-likelihood with respect to Θ and Γ . We introduce two different penalty functions $P_\lambda(\cdot)$ and $P_\rho(\cdot)$ for intra time-slice conditional independencies Θ , and inter time-slice conditional independencies Γ , respectively. Therefore, the objective function for optimization can be defined as

$$Q_{pen}(\Theta, \Gamma | \Theta_\lambda^*, \Gamma_\rho^*) = \frac{n(T-1)}{2} \left[\log \det(\Theta) - \text{tr} \left(\Theta S_\Gamma^{(E)} \right) \right] - \sum_{j \neq j'}^p P_\lambda(|\theta_{jj'}|) - \sum_{j,j'}^p P_\rho(|\gamma_{jj'}|) \quad (5.17)$$

where $S_\Gamma^{(E)}$ denotes the expectation of S_Γ given the data and updated parameters, and $\theta_{jj'}$ and $\gamma_{jj'}$ are the jj' -th element of the Θ and Γ matrices. Among different penalty functions, we consider the L_1 norm and smoothly clipped absolute deviation (SCAD) penalty functions which have the most desirable sparsity properties.

L_1 penalized EM. The Lasso or L_1 penalty function is defined as

$$P_\lambda(\theta) = \lambda|\theta|.$$

The L_1 penalty leads to a desirable optimization problem, where the log-likelihood is convex and can be solved efficiently using various optimization algorithms at the k -th iteration

of the EM. Under this penalty function, the updated estimates are given via

$$(\Theta_\lambda^{(k)}, \Gamma_\rho^{(k)}) = \arg \max_{\Theta, \Gamma} \left\{ \log \det(\Theta) - \text{tr} \left(S_\Gamma^{(E)} \Theta \right) - \lambda \sum_{j \neq j'}^p |\theta_{jj'}| - \rho \sum_{j, j'}^p |\gamma_{jj'}| \right\} \quad (5.18)$$

where the sparsity level of intra and inter time-slice conditional independences are controlled by λ and ρ . L_1 penalty is biased due to its constant rate of penalty. To address this issue, Fan and Li (2001) proposed the SCAD penalty, which results in unbiased estimates for large coefficients.

SCAD penalized EM. The SCAD penalty function is expressed as

$$P_{\lambda, a}(\theta) = \begin{cases} \lambda|\theta| & \text{if } \theta \leq \lambda, \\ -\frac{|\theta^2| - 2a\lambda|\theta| + \lambda^2}{2(a-1)} & \text{if } \lambda \leq |\theta| \leq a\lambda, \\ \frac{(a+1)^2\lambda^2}{2} & \text{if } |\theta| > a\lambda. \end{cases}$$

where λ and a are two tuning parameters. The function $P_{\lambda, a}(\theta)$ corresponds to a quadratic spline on $[0, \infty)$ with knots at λ and $a\lambda$. A similar function can be written for $P_{\rho, a}(\gamma)$ where ρ and $a\rho$ are two knots. The SCAD penalty is symmetric but non-convex; its first order derivative is given by

$$P'_{\lambda, a}(\theta) = \lambda \left\{ I(|\theta| \leq \lambda) + \frac{(a\lambda - |\theta|)_+}{(a-1)\lambda} I(|\theta| > \lambda) \right\}, \quad a > 2$$

The notation z_+ stands for the positive part of z . Fan and Li (2001) showed that in practice $a = 3.7$ is a good choice. Maximizing non-convex penalized likelihood is challenging. To address this issue we use an efficient algorithm proposed in Fan et al. (2009), which is based on local linear approximation, to maximize the penalized log-likelihood for the SCAD penalty function. In each step, a symmetric linear function is used to locally approximate the SCAD penalty. Using the Taylor expansion, $P_{\lambda, a}(\theta)$ and $P_{\rho, a}(\gamma)$ can be approximated in the neighborhood of θ_0 and γ_0 as follows:

$$P_\lambda(|\theta|) \approx P_\lambda(|\theta_0|) + P'_\lambda(|\lambda_0|)(|\theta| - |\theta_0|)$$

$$P_\rho(|\gamma|) \approx P_\rho(|\gamma_0|) + P'_\rho(|\rho|)(|\gamma| - |\gamma_0|).$$

Due to the monotonicity of $P_\lambda(\cdot)$ and $P_\rho(\cdot)$ over $[0, \infty)$, the derivatives $P'_\lambda(\cdot) = \frac{\partial}{\partial \theta}(P_\lambda(\theta))$ and $P'_\rho(\cdot) = \frac{\partial}{\partial \gamma}(P_\rho(\gamma))$ are non-negative for $\theta \in [0, \infty)$ and $\gamma \in [0, \infty)$. Therefore, under the penalized log-likelihood with the SCAD penalty, estimation of the sparse parameters $\Theta^{(k)}$ and $\Gamma^{(k)}$ relies on the solution of the following optimization problem at step k

$$(\Theta_\lambda^{(k)}, \Gamma_\rho^{(k)}) = \arg \max_{\Theta, \Gamma} \left\{ \log \det(\Theta) - \text{tr}(S_\Gamma^{(E)} \Theta) - \sum_{j \neq j'}^p w_{jj'} |\theta_{jj'}| - \sum_{j,l}^p \nu_{jl} |\gamma_{jl}| \right\} \quad (5.19)$$

where $w_{jj'} = P'_\lambda(\theta_{jj'}^{(k)})$, $\nu_{jl} = P'_\rho(\gamma_{jl}^{(k)})$, and $\theta_{jj'}^{(k)}$, $\gamma_{jl}^{(k)}$ are jj' -th element of Θ and jl -th element of Γ , respectively. The SCAD penalty applies a constant penalty to large coefficients, whereas the L_1 penalty increases linearly as $|\theta|$ increases. This feature keeps the SCAD penalty from producing biases to estimate large coefficients. Therefore, the SCAD penalty overcomes the bias issue of the L_1 penalty. Then a two-stage-optimization problem within the M-step of the EM algorithm is employed to solve the objective functions (5.18) or (5.19) to estimate the parameters Θ and Γ .

Glasso calculation of $\Theta^{(k)}$. For the SCAD penalty-based estimation, in the first stage we optimize

$$\Theta_\lambda^{(k)} = \arg \max_{\Theta} \left\{ \log \det(\Theta) - \text{tr}(S_{\Gamma^*}^{(E)} \Theta) - \sum_{j \neq j'}^p w_{jj'} |\theta_{jj'}| \right\},$$

for previous Γ^* . This optimization can be solved efficiently using the graphical lasso algorithm proposed by Friedman et al. (2008). Due to the sparsity in each iteration, we consider a one-step local linear approximation algorithm (LLA). Zou and Li (2008) showed that one-step LLA, asymptotically, performs as well as the fully iterative LLA algorithm as long as the initial solution is good enough. In practice, we take the initial value as the L_1 penalty graphical LASSO to estimate the intra time-slice conditional independences Θ in order to calculate the initial weights $w_{jj'}$ and ν_{jl} .

Regularized coordinate descent algorithm for $\Gamma^{(k)}$. After we finish an updating Θ in the first-stage of the optimization, we proceed in the second-stage to update the estimate of Γ given the updated Θ . In the SCAD penalty-based we optimize

$$\begin{aligned}
\Gamma_\rho^{(k)} &= \arg \max_{\Gamma} \left\{ \log \det(\Theta_\lambda^{(k)}) - \text{tr}(S_\Gamma^{(E)} \Theta_\lambda^{(k)}) - \sum_{j,l}^p \nu_{jl} |\gamma_{jl}| \right\} \\
&= \arg \max_{\Gamma} \left\{ \log \det(\Theta_\lambda^{(k)}) - \text{tr}(S_{cc} \Theta_\lambda^{(k)}) - S_{cp} \Gamma' \Theta_\lambda^{(k)} - \Gamma S'_{cp} \Theta_\lambda^{(k)} + \Gamma S_{pp} \Gamma' \Theta_\lambda^{(k)} - \sum_{j,l}^p \nu_{jl} |\gamma_{jl}| \right\}.
\end{aligned} \tag{5.20}$$

This objective function is quadratic in Γ for given $\Theta_\lambda^{(k)}$. Thus, we use a direct coordinate descent algorithm to calculate $\Gamma_\rho^{(k)}$. So, the derivative of the penalized negative log-likelihood (5.20) with respect to γ_{jl} is

$$\frac{\partial \ell_p}{\partial \gamma_{jl}} = -2e'_j(S'_{cp} \Theta_\lambda^{(k)})e_i + 2e'_j(S_{cc} \Gamma' \Theta_\lambda^{(k)})e_i + \nu_{jl} \text{sgn}(\gamma_{jl}) \tag{5.21}$$

where $\text{sgn}(\cdot)$ is the sign function. These are the Karush–Kuhn–Tucker (KKT) equations defining the solution to the maximization problem. We note that for an arbitrary matrix $A_{p \times p}$, $\partial \text{tr}(\Gamma A) / \partial \gamma_{jl} = a_{lj} = e'_l A e_j$, where e_l and e_j are the corresponding base vectors, each with dimension p . Setting the derivative of negative log-likelihood (5.21) at zero, we get an update for the elements of Γ matrix as follows

$$\gamma_{jl} = \text{sgn}(g_{jl}) \frac{(|g_{jl}| - \nu_{jl})_+}{2(e'_l S_{cc} e_l)(e'_j \Theta_\lambda^{(k)} e_j)}, \tag{5.22}$$

where $g_{jl} = 2\{e'_l(S'_{cp} \Theta_\lambda^{(k)})e_j + (e'_l S_{cc} e_l)(e'_j \Theta_\lambda^{(k)} e_j) \gamma_{jl} - e'_l(S_{cc} \Gamma' \Theta_\lambda^{(k)})e_j\}$, γ_{jl} , and $\Gamma_\rho^{(k)}$ are the estimates in the last step of the iteration inside the optimization (5.22).

Given the two-stage optimization problem inside the M-step, we update the S_Γ matrix in the E-step. This iterative procedure continues until the difference between previous $(\Theta_\lambda^{(k-1)}, \Gamma_\rho^{(k-1)})$ and updated $(\Theta_\lambda^{(k)}, \Gamma_\rho^{(k)})$ becomes smaller than a, user specified, tolerance. Based on our simulation experiments, the EM algorithm converges in a few iterations (at most 5 iterations are needed to reach the convergence). We define the estimate as the stationary point of the EM, $(\widehat{\Theta}_\lambda, \widehat{\Gamma}_\rho) = \lim_{k \rightarrow \infty} (\Theta_\lambda^{(k)}, \Gamma_\rho^{(k)})$.

5.2.5 Selection of tuning parameters

To determine the sparsity of the proposed dynamic chain graph model, parameters λ and ρ have to be tuned. We focus on estimating the sparse intra and inter time-slice conditional

Fixed at t= 5	Performance Θ			Performance Γ		
	F_1 score	SEN	SPE	F_1 score	SEN	SPE
p=10 & n=20						
tsnetwork	0.35	0.35	0.77	0.42	0.43	0.68
SparseTSCGM	0.14	0.14	0.89	0.42	0.67	0.34
SparseTSCGM*	0.20	0.18	0.88	0.40	0.47	0.56
p=10 & n=50						
tsnetwork	0.37	0.37	0.85	0.44	0.43	0.7
SparseTSCGM	0.33	0.45	0.80	0.42	0.65	0.34
SparseTSCGM*	0.31	0.32	0.86	0.42	0.45	0.63
p=50 & n=20						
tsnetwork	0.18	0.12	0.98	0.30	0.30	0.93
SparseTSCGM	0.02	0.03	0.95	0.31	0.54	0.81
SparseTSCGM*	0.00	0.00	1.00	0.31	0.22	0.98
p=50 & n=50						
tsnetwork	0.13	0.08	1.00	0.32	0.24	0.95
SparseTSCGM	0.03	0.03	0.97	0.33	0.55	0.82
SparseTSCGM*	0.07	0.04	1.00	0.28	0.25	0.92

Table 5.1 Performance measure results of simulation study for tsnetwork and SparseTSCGM using SCAD penalized likelihood estimation for precision and autoregressive coefficient matrices for fixed time point, t=5. In SparseTSCGM* normal transformation is applied to simulated ordinal data.

independences Θ and Γ , and we employ the Bayesian information criteria (BIC)

$$\begin{aligned}
\text{BIC}(\lambda, \rho) &= -2\ell_Y(\hat{\Theta}_\lambda, \hat{\Gamma}_\rho) + \log(n(T-1)) \left(\text{df}(\hat{\Theta}_\lambda)/2 + \text{df}(\hat{\Gamma}_\rho) + p \right) \\
&\approx n(T-1) \left\{ \log(\det(\hat{\Theta}_\lambda)) - \text{tr}(S_{\hat{\Gamma}_\rho}^{(E)} \hat{\Theta}_\lambda) \right\} \\
&\quad + \log(n(T-1)) \left(\text{df}(\hat{\Theta}_\lambda)/2 + \text{df}(\hat{\Gamma}_\rho) + p \right)
\end{aligned} \tag{5.23}$$

to select tuning parameters λ and ρ , where T and p are the number of time points and the number of variables, respectively; $\text{df}(\hat{\Theta}_\lambda)$ shows the number of non-zero elements in the off-diagonal of $\hat{\Theta}_\lambda$, and $\text{df}(\hat{\Gamma}_\rho)$ is the number of non-zero elements of $\hat{\Gamma}_\rho$. The approximation made in BIC is the result of a Laplace-type of approximation, which makes fast calculation feasible. We choose the optimal value of the penalty parameters, which minimizes $\text{BIC}(\lambda, \rho)$ on a grid of candidate values for λ and ρ . One may consider other information criteria suitable for graph estimations. Wang et al. (2007) and Yin and Li (2011) have shown that BIC performs well for selecting the tuning parameter of penalized likelihood

estimation.

5.3 Simulation study

To investigate and assess the performance of the proposed dynamic chain graph model, we set up a simulation to generate sparse Θ and Γ matrices, similar to Abegaz and Wit (2013), and Yin and Li (2011). Here we evaluate the performance of the proposed method with respect to different random graph structures for Θ and Γ matrices. Different graph structures for Θ can be simulated through the R package *flare*. To generate the Γ matrix we took the upper diagonal of an independently generated Θ along with a 0.2% nonzero diagonal elements sampled from uniform $(0, 1)$, similar to the R package *SparseTSCGM*.

First we simulate data from $N_p(0, \Theta^{-1})$ at time $t = 1$; for the next time steps $t = 2, \dots, T$ we use VAR(1) model such that $Z^{(t)}|Z^{(t-1)} \sim N(\Gamma Z^{(t-1)}, \Theta^{-1})$. Then, n i.i.d samples is generated for each time point. This results in p -variate time series data. Finally, we discretize the obtained time series data with Gaussian marginals into randomized quantile ranges and treat them as categorical time series data. The simulations are repeated 50 times, independently, for different values of p, n, t .

To assess the performance of our proposed method in recovering the intra and inter conditional independence relationships we compute the F_1 -score, sensitivity and specificity measures, which are defined as:

$$F_1 - \text{score} = \frac{2TP}{2TP + FP + FN}, \quad SEN = \frac{TP}{(TP + FN)}, \quad SPE = \frac{TN}{TN + FP}$$

where TP, TN, FP, and FN are the numbers of the true positive, true negative, and false positive, false negative in identifying the non-zero elements in the Θ and Γ matrices. We note that high values of F_1 -score, sensitivity and specificity indicate good performance of a method for the given combination of p, n and t . However, as there is a natural trade-off between sensitivity and specificity, to evaluate the performance of each method we focus particularly on the F_1 -score.

We compare the finite sample performance of the proposed approach using SCAD penalized maximum likelihood with a recently proposed approach implemented in R package *SparseTSCGM* (Abegaz et al., 2015). For further comparison we have applied *SparseTSCGM* to the original simulated ordinal data and to the transferred data using normal transformation. We present the simulation results of sparse precision and autoregressive coefficient matrices in Tables 5.1 and 5.2, based on optimal tuning parameters chosen by the minimum

Fixed at t= 10	Performance Θ			Performance Γ		
	F_1 score	SEN	SPE	F_1 score	SEN	SPE
p=10 & n=20						
tsnetwork	0.35	0.35	0.77	0.43	0.43	0.68
SparseTSCGM	0.23	0.32	0.76	0.40	0.61	0.34
SparseTSCGM*	0.26	0.27	0.88	0.41	0.46	0.57
p=10 & n=50						
tsnetwork	0.38	0.37	0.85	0.44	0.43	0.7
SparseTSCGM	0.40	0.59	0.69	0.41	0.64	0.32
SparseTSCGM*	0.36	0.40	0.86	0.43	0.47	0.61
p=50 & n=20						
tsnetwork	0.11	0.07	0.99	0.31	0.26	0.95
SparseTSCGM	0.02	0.02	0.98	0.33	0.55	0.77
SparseTSCGM*	0.05	0.03	1.00	0.29	0.25	0.93
p=50 & n=50						
tsnetwork	0.37	0.30	0.98	0.31	0.25	0.95
SparseTSCGM	0.39	0.34	0.99	0.24	0.67	0.64
SparseTSCGM*	0.34	0.35	0.97	0.28	0.26	0.92

Table 5.2 Performance measure results of simulation study for tsnetwork and SparseTSCGM using SCAD penalized likelihood estimation for precision and autoregressive coefficient matrices for fixed time point, t=10. In SparseTSCGM* normal transformation is applied to simulated ordinal data.

EBICs. In each simulation setting we have very sparse matrices with only $(1/p) \times 100$ nonzero entries. From the tables we can see that in most cases our method, compared with the alternative method, scores better in terms of the F1-score. These results suggest that, although recovering sparse network structure in ordinal time series data is a challenging task, the proposed approach performs well on model-based simulations. We note here that improved model performance can be gained by allowing the tuning parameters ρ and λ to vary with each simulation.

5.4 Netherlands Study of Depression and Anxiety

We applied our method to a Netherlands Study of Depression and Anxiety (NESDA) Severity of Depression dataset. Depression and anxiety disorders are common at all ages. Approximately one out of three people in the Netherlands will be faced with one of these disorders at some time during their lives. It is still not clear why some people recover quickly and others suffer for long periods of time. The Netherlands Study of Depression

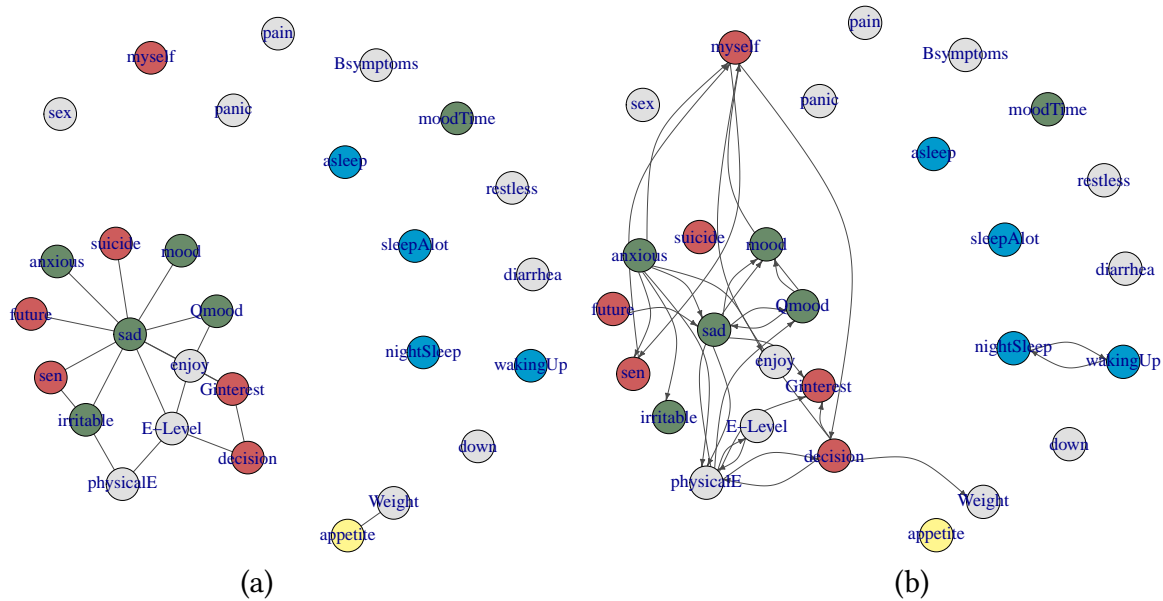


Fig. 5.1 Intra time-slice conditional independence undirected network in NESDA dataset (a) and delayed interactions between items in NESDA across time steps (b). NESDA data are in five categories: (i) sleep, blue, (ii) mood, green, (iii) appetite, yellow, (iv) somatic, gray, (v) mental, red.

and Anxiety (NESDA) was therefore designed to investigate the course of depression and anxiety disorders over a period of several years. The main aim of NESDA is to determine the (psychological, social, biological and genetic) factors that influence the development and long-term prognosis of anxiety and depression. The data consist of 28 items (variables) collected over 3 time intervals. For each of the 28 variables there are four corresponding answers 0=None, 1=Mild, 2=Moderate, 3=Severe. For example, for the item “Feeling sad” there are four corresponding answers, from “0” indicating no depression (e.g., “I do not feel sad”), to “3”, referring to a more severe depressive symptom (e.g., “I feel sad nearly all the time”). A total score is derived (possible range: 0–84), and higher scores indicate relatively severe depressive symptomatology. From the 1799 participants, we selected 200 more informative patients. The BIC criterion selects the penalty values $\lambda = 0.19$ and $\rho = 0.23$. The resulting instantaneous and delayed interaction networks among the 28 items are shown in Figure 5.1, left and right panels respectively.

Figure 5.1(a) shows undirected links that suggest contemporaneous interactions among 12 items, and Figure 5.1(b) displays directed edges that indicate Granger-causality relationships or delayed interactions between these 12 items. We observe that item “Feeling sad” is the hub in both figures, suggesting that it plays a fundamental role in treating depression and anxiety disorders. Also, Figure 5.1(b) shows that there are several directed links pointing from mood category to mental category; this suggests that mood disorders influ-

ence the long-term development of mental disorders. Interestingly, Figure 5.1b shows that sleeping disorders have no effect on other symptoms of depression.

5.5 Discussion

We have presented a dynamic model for multivariate ordinal time series data which assumes a chain graph representation of the conditional independence structure among time series components. The proposed model combines Gaussian copula graphical models and dynamic Bayesian networks to infer instantaneous conditional dependence relationships among time series components and dynamic or delayed interactions, possibly potentially “causal” relationships, among variables at consecutive time steps. The directed edges reflect Granger causality whereas the undirected edges represent the contemporaneous dependence structure.

To obtain sparse estimates for the instantaneous conditional dependence graph and for the Granger-causality graph, we considered penalized log-likelihood estimation using the L_1 and SCAD penalties. Simulation studies show that the proposed sparse estimates reflect the underlying intra- and inter-time slice conditional dependence networks more accurately compared to the alternative method.

The method was applied to the Netherlands study of depression and anxiety categorical time series data. The model has, however, much wider applicability to any multivariate mixed continuous-and-discrete time series data.

5.6 Appendix

Applying Lemma 5.1 to the conditional expectations in (5.15) and (5.16) leads to the following approximations:

$$\begin{aligned}
 E(Z_{i,j}^{(t)} | y_i; \Theta^*, \Gamma^*) &\approx \Sigma_{j,-j} \Sigma_{-j,-j}^{-1} \left(E(Z_{i,-j}^{(t)\tau} | y_i; \Theta^*, \Gamma^*) - (\Gamma_i E(Z_i^{(t-1)} | y_i; \Theta^*, \Gamma^*))_{-j} \right) \\
 &\quad + (\Gamma_i E(Z_i^{(t-1)} | y_i; \Theta^*, \Gamma^*))_j + \frac{\phi(\delta_{i,j,y_{i,j}^{(t)}}^{(t)} - \phi(\delta_{i,j,y_{i,j}^{(t)+1}}^{(t)})}{\Phi(\delta_{i,j,y_{i,j}^{(t)+1}}^{(t)}) - \Phi(\delta_{i,j,y_{i,j}^{(t)}}^{(t)})} \sigma_{i,j} \quad (5.24)
 \end{aligned}$$

$$\begin{aligned}
 E((Z_{i,j}^{(t)})^2 | y_i; \Theta^*, \Gamma^*) &\approx \left((\Gamma_i E(Z_i^{(t-1)} | y_i; \Theta^*, \Gamma^*))_j \right)^2 + \Sigma_{j,-j} \Sigma_{-j,-j}^{-1} E(Z_{i,-j}^{(t)\tau} Z_{i,-j}^{(t)} | y_i; \Theta^*, \Gamma^*) \Sigma_{-j,-j}^{-1} \\
 &\quad \Sigma_{-j,j} + \Sigma_{j,-j} \Sigma_{-j,-j}^{-1} \left((\Gamma_i E(Z_i^{(t-1)} | y_i; \Theta^*, \Gamma^*))_{-j} \right)^2 \Sigma_{-j,-j}^{-1} \Sigma_{-j,j} - 2\Sigma_{j,-j}
 \end{aligned}$$

$$\begin{aligned}
& \Sigma_{-j,-j}^{-1} E(Z_{i,-j}^{(t)\tau} | y_i; \Theta^*, \Gamma^*) \left((\Gamma_i E(Z_i^{(t-1)} | y_i; \Theta^*, \Gamma^*)) \right)_{-j} \Sigma_{-j,-j}^{-1} \Sigma_{-j,j} \\
& + 2 \left(\Gamma_i E(Z_i^{(t-1)} | y_i; \Theta^*, \Gamma^*) \right)_j \Sigma_{j,-j} \Sigma_{-j,-j}^{(-1)} E(Z_{i,-j}^{(t)} | y_i; \Theta^*, \Gamma^*) \\
& - 2 (\Gamma_i E(Z_i^{(t-1)} | y_i; \Theta^*, \Gamma^*))_j \Sigma_{j,-j} \Sigma_{-j,-j}^{(-1)} \left(\Gamma_i E(Z_i^{(t-1)} | y_i; \Theta^*, \Gamma^*) \right)_{-j} + \sigma_{i,j}^2 \\
& + 2 \frac{\phi(\delta_{i,j,y_{i,j}^{(t)}}^{(t)}) - \phi(\delta_{i,j,y_{i,j}^{(t)+1}}^{(t)})}{\Phi(\delta_{i,j,y_{i,j}^{(t)+1}}^{(t)}) - \Phi(\delta_{i,j,y_{i,j}^{(t)}}^{(t)})} \left[(\Gamma_i E(Z_i^{(t-1)} | y_i; \Theta^*, \Gamma^*))_j \right. \\
& \left. + \Sigma_{j,-j} \Sigma_{-j,-j}^{-1} \left(E(Z_{i,-j}^{(t)\tau} | y_i; \Theta^*, \Gamma^*) - (\Gamma_i E(Z_i^{(t-1)} | y_i; \Theta^*, \Gamma^*))_{-j} \right) \right] \sigma_{i,j} \\
& + \frac{\delta_{i,j,y_{i,j}^{(t)}}^{(t)} \phi(\delta_{i,j,y_{i,j}^{(t)}}^{(t)}) - \delta_{i,j,y_{i,j}^{(t)+1}}^{(t)} \phi(\delta_{i,j,y_{i,j}^{(t)+1}}^{(t)})}{\Phi(\delta_{i,j,y_{i,j}^{(t)+1}}^{(t)}) - \Phi(\delta_{i,j,y_{i,j}^{(t)}}^{(t)})} \sigma_{i,j}^2 \tag{5.25}
\end{aligned}$$

where $\delta_{i,j,y_{i,j}^{(t)}}^{(t)} = (c_{i,j}^{(t)} - \mu'_{i,j}) / \sigma'_{i,j}$. Here, the first order delta method is used to approximate the nonlinear terms (more details in (Guo et al., 2015)).

Another approximation that can be replaced in (5.15) and (5.16) is as follows:

$$E\left(Z_{i,j}^{(t)} | y_i; \Theta^*, \Gamma^*\right) = E\left[E\left(Z_{i,j}^{(t)} | Z_{i,-j}^{(t)}, y_{i,j}^{(t)}; \Theta, \Gamma\right) | y_i; \Theta^*, \Gamma^*\right] \tag{5.26}$$

$$E\left(Z_{i,j}^{(t)2} | y_i; \Theta^*, \Gamma^*\right) = E\left[E\left(Z_{i,j}^{(t)2} | Z_{i,-j}^{(t)}, y_{i,j}^{(t)}; \Theta, \Gamma\right) | y_i; \Theta^*, \Gamma^*\right] \tag{5.27}$$

where $Z_{i,-j}^{(t)}$ represents a set that contains all the variables at time step t , except the j -th variable.

Within each time step, the mean $\mu_{i,j}$ is a linear function of $z_{i,-j}^{(t)}$, and both $\frac{\phi(\delta_1) - \phi(\delta_2)}{\Phi(\delta_2) - \Phi(\delta_1)}$ and $\frac{\delta_1 \phi(\delta_1) - \delta_2 \phi(\delta_2)}{\Phi(\delta_2) - \Phi(\delta_1)}$ are nonlinear functions of $z_{i,-j}^{(t)}$. Applying Lemma 5.1 to the conditional expectations in (5.26) and (5.27) leads to the following approximations:

$$E(Z_{i,j}^{(t)} | y_i^{(t)}; \Theta^*, \Gamma^*) \approx \Sigma_{j,-j} \Sigma_{-j,-j}^{-1} E(Z_{i,-j}^{(t)'} | y_i^{(t)}; \Theta^*, \Gamma^*) + \frac{\phi(\delta_{i,j,y_{i,j}^{(t)}}^{(t)}) - \phi(\delta_{i,j,y_{i,j}^{(t)+1}}^{(t)})}{\Phi(\delta_{i,j,y_{i,j}^{(t)+1}}^{(t)}) - \Phi(\delta_{i,j,y_{i,j}^{(t)}}^{(t)})} \sigma_j^{(i)}, \tag{5.28}$$

$$\begin{aligned}
E((Z_{i,j}^{(t)2}) | y_i^{(t)}; \Theta^*, \Gamma^*) & \approx \Sigma_{j,-j} \Sigma_{-j,-j}^{-1} E(Z_{i,-j}^{(t)'} Z_{i,-j}^{(t)} | y_i^{(t)}; \Theta^*, \Gamma^*) \Sigma_{-j,-j}^{-1} \Sigma'_{j,-j} + \sigma_{i,j}^2 \\
& + 2 \frac{\phi(\delta_{i,j,y_{i,j}^{(t)}}^{(t)}) - \phi(\delta_{i,j,y_{i,j}^{(t)+1}}^{(t)})}{\Phi(\delta_{i,j,y_{i,j}^{(t)+1}}^{(t)}) - \Phi(\delta_{i,j,y_{i,j}^{(t)}}^{(t)})} [\Sigma_{j,-j} \Sigma_{-j,-j}^{-1} E(Z_{i,-j}^{(t)\tau} | y_i^{(t)}; \Theta^*, \Gamma^*)] \tilde{\sigma}_{i,j} \\
& + \frac{\delta_{i,j,y_{i,j}^{(t)}}^{(t)} \phi(\delta_{i,j,y_{i,j}^{(t)}}^{(t)}) - \delta_{i,j,y_{i,j}^{(t)+1}}^{(t)} \phi(\delta_{i,j,y_{i,j}^{(t)+1}}^{(t)})}{\Phi(\delta_{i,j,y_{i,j}^{(t)+1}}^{(t)}) - \Phi(\delta_{i,j,y_{i,j}^{(t)}}^{(t)})} \sigma_{i,j}^2, \tag{5.29}
\end{aligned}$$

where $\delta_{i,j,y_{i,j}}^{(t)} = (c_{i,j}^{(t)} - \mu_{i,j})/\sigma_{ij}$. Here, the first order delta method is used to approximate the nonlinear terms. Moreover, we approximate the elements of conditional expectation matrices S_{pp} , S_{cc} , and S_{cp} through equations (5.28) and (5.29).

References

- Abegaz, F. and E. Wit (2013). Sparse time series chain graphical models for reconstructing genetic networks. *Biostatistics*, kxt005.
- Abegaz, F., E. Wit, and M. F. Abegaz (2015). Package ‘sparsetscgm’.
- Behrouzi, P. and E. Wit (2017). Detecting epistatic selection with partially observed genotype data using copula graphical models. *arXiv preprint arXiv:1710.00894*.
- Chandler, D. (1987). Introduction to modern statistical mechanics. *Introduction to Modern Statistical Mechanics, by David Chandler, pp. 288. Foreword by David Chandler. Oxford University Press, Sep 1987. ISBN-10: 0195042778. ISBN-13: 9780195042771, 288.*
- Colombo, D., M. H. Maathuis, M. Kalisch, and T. S. Richardson (2012). Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, 294–321.
- Dahlhaus, R. and M. Eichler (2003). Causality and graphical models in time series analysis. *Oxford Statistical Science Series*, 115–137.
- Dobra, A., A. Lenkoski, et al. (2011). Copula gaussian graphical models and their application to modeling functional disability data. *The Annals of Applied Statistics* 5(2A), 969–993.
- Fan, J., Y. Feng, and Y. Wu (2009). Network exploration via the adaptive lasso and scad penalties. *The annals of applied statistics* 3(2), 521.
- Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association* 96(456), 1348–1360.
- Friedman, J., T. Hastie, and R. Tibshirani (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9(3), 432–441.

- Gao, W. and Z. Tian (2010). Latent ancestral graph of structure vector autoregressive models. *Journal of Systems Engineering and Electronics* 21(2), 233–238.
- Genest, C., K. Ghoudi, and L.-P. Rivest (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 543–552.
- Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 424–438.
- Green, P. J. (1990). On use of the em for penalized likelihood estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 443–452.
- Guo, J., E. Levina, G. Michailidis, and J. Zhu (2015). Graphical models for ordinal data. *Journal of Computational and Graphical Statistics* 24(1), 183–204.
- Hoff, P. D. (2007). Extending the rank likelihood for semiparametric copula estimation. *The Annals of Applied Statistics*, 265–283.
- Johnson, N., S. Kotz, and N. Balakrishnam (1995). Noncentral χ^2 distributions. noncentral f distributions. *Continuous univariate distributions* 2, 433.
- Kalisch, M. and P. Bühlmann (2007). Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research* 8(Mar), 613–636.
- Lauritzen, S. L. (1996). *Graphical models*, Volume 17. Clarendon Press.
- Lauritzen, S. L. and N. Wermuth (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *The annals of Statistics*, 31–57.
- Mohammadi, A., E. C. Wit, et al. (2015). Bayesian structure learning in sparse gaussian graphical models. *Bayesian Analysis* 10(1), 109–138.
- Wang, H., R. Li, and C.-L. Tsai (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* 94(3), 553.
- Yin, J. and H. Li (2011). A sparse conditional gaussian graphical model for analysis of genetical genomics data. *The annals of applied statistics* 5(4), 2630.
- Zou, H. and R. Li (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of statistics* 36(4), 1509.

