Constructing graphical models via the focused information criterion

Claeskens, G.; Pircalabelu, E.; Waldorp, L.

DOI
10.2139/ssrn.2419382
Publication date
2014
Document Version
Submitted manuscript

KU LEUVEN

FACULTY OF ECONOMICS AND BUSINESS

Constructing graphical models via the focused information criterion

Claeskens G, Pircalabelu E, Waldorp L.

KBI_1404

Constructing Graphical Models via the Focused Information Criterion

Gerda Claeskens, Eugen Pircalabelu and Lourens Waldorp

Abstract A focused information criterion is developed to estimate undirected graphical models where for each node in the graph a generalized linear model is put forward conditioned upon the other nodes in the graph. The proposed method selects a graph with a small estimated mean squared error for a user-specified focus, which is a function of the parameters in the generalized linear models, by selecting an appropriate model at each node. For situations where the number of nodes is large in comparison with the number of cases, the procedure performs penalized estimation with quadratic approximations to several popular penalties. To show the procedure's applicability and usefulness we have applied it to two datasets involving voting behavior of U.S. senators and to a clinical dataset on psychopathology.

1 Introduction

We propose a focused search method of estimating an undirected graph when the distribution of the random variables associated with each node is a member of an exponential family of distributions, including the Gaussian, Poisson and binomial distributions as special cases. The graph is constructed nodewise, hence instead of solving one multivariate optimization problem which in this case is difficult in gen-

Gerda Claeskens
KU Leuven, ORSTAT and Leuven Statistics Research Center, Naamsestraat 69, 3000 Leuven, Belgium, e-mail: gerda.claeskens@kuleuven.be

Eugen Pircalabelu
KU Leuven, ORSTAT and Leuven Statistics Research Center, Naamsestraat 69, 3000 Leuven, Belgium, e-mail: eugen.pircalabelu@kuleuven.be

Lourens Waldorp
University of Amsterdam, Department of Psychological Methods, Weesperplein 4, 1018 Amsterdam, The Netherlands, e-mail: waldorp@uva.nl

eral, we proceed by optimizing many univariate problems (one for each node) and then 'glue together' all the pieces of information.

By the *focus* of the research we mean a predefined function of the model parameters, such as the mean of a response variable in a regression model. This focus we wish to estimate well in the sense of having a low mean squared error (MSE). In the nodewise approach we fit at each node a generalized linear model (GLM), implying that selecting the neighboring nodes, or equivalent, selecting the edges in the graph, is nothing but a variable selection problem in a generalized linear model. Obviously, different models give rise to different bias and variance quantities for the focus estimator, and thus searching a model which produces a small MSE for an estimator is a sensible thing to do. Moreover, a researcher can have different focuses which reflect his/her scientific interests and thus one can estimate using a given dataset several (possibly different) graphs which serve the corresponding research purposes. With this in mind, we point out that the focused approach may take more carefully the domain knowledge into account that is available to the researcher when defining the focus, and outputs a model fine-tuned for that specific focus. Thus this approach moves from a 'one model for all purposes' scheme, to a 'one model per purpose' approach.

Graphical models visualize relations that exist between components of a multivariate random vector, say $X = (X_1, \ldots, X_p)$. In a graph, each component of this vector is identified with a node, and a relation between two components is visualized by drawing a connection, an edge, between the corresponding nodes. For an example, see Figure 3. Different types of relations between the random components may be represented by different types of edges (with or without arrowheads).

The most common types of graphical models to be encountered in the literature are Bayesian networks and Markov networks. In terms of graphical representations the Bayesian networks are based on directed graphs (lines with arrowheads) while Markov networks are based on undirected graphs. Both types of models try to graphically encode the conditional independencies that hold between variables that are represented here by nodes in the graph. In the case of directed graphs drawing a directed edge as $i \rightarrow j$ is to be understood that node i influences node j or that node i 'causes' in some sense node j. For example one might represent a relation between Age and Income in a graph as $Age \rightarrow Income$ with a clear message that one's income depends on ones age as on average older people earn more than younger, but refute the relation $Age \leftarrow Income$ as this makes probably no 'causal' sense. On the other hand if one faces a situation where a causal effect cannot be assumed in any of the two directions then one can find it useful to place undirected edges between the two nodes, in order to signalize that there exists an association between these two nodes though without a precise directionality effect. In a genomics study one might assume that gene i is correlated with gene j, and represent it by an undirected edge as it might be implausible that any of the genes has a direct effect on the other. For a comprehensive explanation about graphical models, we refer to Lauritzen (1996) or Cox and Wermuth (1996).

We here concentrate on undirected graphical models. For Gaussian random vectors, having an edge in an undirected graph yields the following interpretation. Ran-

dom variable X_i is dependent on X_j conditioned on all remaining variables in the multivariate vector, if and only if there is an edge in the graph between the nodes representing these variables X_i and X_j. Equivalently, there is a zero entry at the crossing of row i and column j in the inverse covariance matrix, also called the concentration matrix, if and only if no edge is drawn between the corresponding nodes i and j in the graph. Thus, estimating a graphical model in the sense of drawing edges between nodes, is equivalent to determining the positions of the zero and non-zero entries in the concentration matrix, see Dempster (1972) and Lauritzen (1996). Thus, in case one discovers an entry in the concentration matrix that is zero, or equivalently, one finds conditional independencies, there is a simpler way of writing the joint distribution of the multivariate vector X, that adequately describes the relations between the components of X.

Let us consider a sample of n multivariate random vectors $X_k = (X_{k1}, \ldots, X_{kp}), k = 1, \ldots, n$, each consisting of p components. One way to estimate the non-zero entries in the concentration matrix is through nodewise regression models (Meinshausen and Bühlmann, 2006). In turn, each random variable associated with a single node (say node i) is taken as the response variable and the variables corresponding to the other nodes act as covariates (predictors). A non-zero entry is considered to exist at row i and column j ($i, j = 1, \ldots, p$) when, for Gaussian data, the coefficient $\beta_{ij} \neq 0$ in the regression model with the variable corresponding to node i as the response

$$X_{ki} = \beta_{i0} + \sum_{l=1,\ldots,p, l\neq i} \beta_{il} X_{kl} + \varepsilon_{ki}, \tag{1}$$

and at the same time, $\beta_{ji} \neq 0$ in the regression model with the variable corresponding to node j as the response variable

$$X_{kj} = \beta_{j0} + \sum_{l=1,\ldots,p, l\neq j} \beta_{jl} X_{kl} + \varepsilon_{kj}, \tag{2}$$

with ε_{ki} and ε_{kj} independent normal random variables with zero mean and $k = 1, \ldots, n$ observations. This is referred to as an 'AND' rule. One could also use an 'OR' rule that includes an edge between nodes i and j when either β_{ij} '*or*' β_{ji} is nonzero (we refer to Meinshausen and Bühlmann, 2006, for an application based on the two rules). Throughout the paper the 'OR' rule is applied for constructing the graphs, due to the high-dimensionality of the problem and the greedy manner in which nodewise models are constructed. The 'OR' rule might overfit by including spurious edges, but one would rather have a model that overfits (i.e. not missing some important edges) than a model that underfits (missing important edges).

In Pircalabelu et al. (2012) we propose to use the focused information criterion (FIC, Claeskens and Hjort, 2003), which is driven by the mean squared error, to select the variables in the above nodewise regression models. Once the neighbors for all variables in the nodewise regression models are selected by FIC, we can draw the selected graph. This is referred to as the *FIC selected graph*. See Section 4.2. This idea is extended to larger graphs in Pircalabelu et al. (2013), using penalized estimation methods, see also Section 4.3.

A main reason for using the focused information criterion and not any other variable selection method, is that this criterion makes it possible to obtain tailor-made graphs. For instance, graphs representing interconnectivity in mental symptoms (see e.g., Borsboom et al., 2011) can provide predictions of the development of a disorder in patients. Such predictions are optimal whenever the graph used to represent the disorder is tuned to certain types of predictions (see the example in Section 2 on psychopathology for more details). In a statistical sense, a good estimator is one with a low mean squared error (MSE), which is defined as the sum of the squared bias and the variance of the estimator.

For each node as a 'response variable' we have for each remaining variable the choice to include it or not to include it as a covariate, resulting in a list of possible models. In each such model we can estimate the focus. Underlying the FIC are estimators of the mean squared error of the focus estimators in each of the different regression models under consideration. Minimizing the FIC is equivalent to minimizing the estimated MSE of the focus over the different models. Thus, we select for each node a regression model and use this in a next step to construct a graph, that is aimed to give a low MSE for the estimated focus.

The tailor-made aspect of FIC is easily understood. Specifying a different focus, will result in different focus estimators and thus in different MSE values, and consequently different FIC values. Hence, different focuses may lead to different graphs. Each time, we select that graph that scores best in estimated MSE (that is, FIC) for that focus. More details are given in Section 4.

In this chapter we extend the methodology of the FIC for graphs based on Gaussian random variables, to graphs for multivariate random vectors where the nodes may be fit through generalized linear models (McCullagh and Nelder, 1989).

2 Data examples

2.1 Data example on 'Dynamics of psychopathology'

The data used in this subsection come from a study of van Borkulo et al. (2013) and consist of a series of measurements for two subjects: a rapid cycling bipolar patient and a healthy control case. A bipolar patient has episodes of mania (energetic, highly productive, etc.) and/or depression; on average 0.5 episodes per year. A rapid cycling bipolar patient has at least four such episodes per year. Both subjects were asked to rate their feelings during 93 days on the Positive and Negative Affect scale (PANAS, Watson et al., 1988). The scale consists of 22 feelings or emotions and during each day the two subjects were asked to rate on a 5 point Likert scale (ranging from 'not at all' to 'extremely') to what extent the feeling pertains to them. All variables have been discretized to 0/1 binary values where 0 indicated 'not at all', while 1 indicated all other categories. Afterwards, positive affect items were reversed scored, such that for a positive affect item a '0' value represents the presence

of the positive feeling while on the negative affect item the same value represents the absence of a negative value. The purpose of the recoding was to concentrate on subjects that have had positive feelings compared to subjects that lacked to have these feelings. For example, it means that if for a subject the value 0 is recorded for 'feeling interested' and 0 is recorded for 'feeling distressed and unhappy' then the subject is likely to have felt interested but *not* distressed and unhappy.

All 22 feelings were considered as nodes in a network that influence each other.

The main goal of those authors was to numerically quantify differences between the patient and control in the contact process framework (see Fiocco and van Zwet, 2004). In the contact process an infected node (determined by a value of 1) at time t can infect its immediate neighbors, which in turn can infect their other immediate neighbors. As time passes some of the previously infected nodes can also recover (switching from 1 to 0). Two independent Poisson processes are assumed: spontaneous recovery of infected nodes (with rate μ) and infection of healthy nodes (with rate proportional to λ). The estimated ratio $\rho = \lambda/\mu$, called the 'basic reproduction number' (BRP) is then used to quantify the differences between the two subjects. The analysis in van Borkulo et al. (2013) suggests that for the bipolar patient the BRP is much higher than that of the control, meaning that for the patient the network will continue to be infected indefinitely.

One of the main assumptions of the model is that the researcher has a network at his/her disposal on which the infections and recoveries can be observed, and as such we wish to put forward possible networks after which, based on the estimated networks, we will estimate and compare the BRP for both subjects.

2.2 Data example on U.S. voting behavior

The data set used here encodes the U.S. senate voting records data from the 109th congress between 2004 and 2006 (see Banerjee et al., 2008). It contains only binary 0/1 variables where a '0' represents a 'No' vote for the proposed bill and a '1' marks a 'Yes' vote. There are 100 variables, corresponding to 100 senators (64 of them being Democrats and 36 being Republican) and 542 cases, corresponding to 542 bills and amendments put to vote. As in the original paper, all missing votes per bill have been recoded as 'No' votes. The aim of the analysis is to estimate an undirected graph structure where each node represents a senator and each edge between two nodes represents a form of interaction between senators such that the voting behavior of one senator could be used as a predictor for the behavior of another senator. The entire dataset corresponding to 100 senators and 542 bills has been used in the analysis.

We are interested in the describing how the voting behavior of the senators depends on the voting behavior of all other senators. Therefore, we use as a focus the expected value of a node conditioned on the values of all other nodes. To this end we will use the voting pattern of all senators for the 'Flag Desecration' amendment sometimes referred to in the media as the 'flag-burning' amendment. The initiative

proposed a constitutional amendment that would allow the U.S. Congress to outlaw the physical desecration of the flag of the United States. A vivid debate was started between supporters of the freedom of speech and supporters of national symbols, and the attempt to adopt such an amendment failed by only one vote. All senators have given their vote on the bill and there was no missing information for this focus. We wish to estimate the undirected graphical structure that provides the smallest MSE of the focus estimator at each node, using the procedure described in Section 4. Since there are 100 nodes in this example, we will use the penalized approaches of Section 4.3.

2.3 Data example on hunting spider species

The data come from a study of van der Aart and Smeenk-Enserink (1975) and consists of abundances (numbers trapped over a 60 week period) of hunting spiders in a Dutch dune area. There were 28 sites where data on 12 spider species were collected. In addition, the dataset contains measurements on 6 extra environmental variables for each studied area. The interest here lies in knowing whether and how selected graphs differ for two locations from the dataset. It is of interest to know whether environmental characteristics influence the structure of the estimated networks, as it is expected that some species might prefer to inhabit one type of environment while others might be less influenced by area characteristics. For this purpose we use the observed counts for each species at two locations for which the amount of fallen leaves, moss or the herb layer and the reflection of the soil surface are quite different (see Figure4). The hypothesis is that if the abundance of spiders was not related to area characteristics, the spider counts would be similar at the two locations and the estimated graphical models for these two focuses would be quite similar. Differences between the two estimated graphs can thus be linked to the effects area characteristics have on the presence of spiders.

3 Generalized linear models and graphs

A p-variate random variable $X = (X_1, \ldots, X_p)$ may be represented by a *graph* $\mathscr{G}$. A graph is mathematically defined by a pair of sets $(\mathscr{E}, \mathscr{V})$ where $\mathscr{V}$ is the set of nodes $\{1, \ldots, p\}$, each node j is identified with a univariate variable X_j, $j = 1, \ldots, p$, and where the set of edges $\mathscr{E}$ is a subset of $\mathscr{V} \times \mathscr{V}$ consisting of pairs of distinct nodes.

While in a Gaussian graph X follows a multivariate normal distribution, other distributions may be assumed. We here consider the situation that each component of X has a distribution belonging to an exponential family, such that we may fit nodewise generalized linear models, extending upon the linear models as in (1).

In a generalized linear model, the response Y has a distribution of the type

$$f(y;\vartheta,\phi)=\exp\Big\{\frac{y\vartheta-b(\vartheta)}{a(\phi)}+c(y,\phi)\Big\}, \tag{3}$$

where ϑ and ϕ are unknown parameters and where the functions a, b and c are known. The parameter ϕ is a scale parameter, and ϑ is the main parameter of interest, since it holds that $\mathrm{E}(Y)=\partial b(\vartheta)/\partial\vartheta=b'(\vartheta)$. Another interesting aspect of such a distribution is that $\mathrm{Var}(Y)=a(\phi)\partial^2 b(\vartheta)/\partial\vartheta_i^2$ (see e.g., McCullagh and Nelder, 1989).

Common examples of this exponential family include the normal, Poisson, binomial and gamma distributions. For regression models where each observation may comprise of a vector of covariate values, the parameter ϑ may be taken differently for each observation.

While in a linear model the mean of the response $\mathrm{E}(Y|x)=x^t\beta$ is a linear function, in a generalized linear model there is a monotone and smooth link function denoted by g such that $\mathrm{g}\{\mathrm{E}(Y|x)\}=x^t\beta$. The special choice of $\mathrm{g}(\cdot)=(b')^{-1}(\cdot)$ is referred to as the canonical link. For Bernoulli distributions the logistic link is canonical, the identity function is canonical for normal distributions and for Poisson data it is the log-function.

While it would lead too far to construct a complete list of the existing work on Gaussian and 0/1 binary data for graph construction, we refer to some recent work of Yang et al. (2012), Lee and Hastie (2012), Jalali et al. (2010) and Loh and Wainwright (2012) who construct procedures oriented towards either situations where X is a discrete random variable, or situations where the distribution of X is a member of the more general exponential family of distributions. The above mentioned works are relevant to our case for several reasons. First, they work nodewise, where models are first selected at the level of the nodes and then everything is 'glued' together, and more importantly they also suggest that such nodewise constructions have merit because under certain conditions they are able to recover aspects of the true underlying graph.

Starting from a general form of a univariate exponential distribution, Yang et al. (2012) formulate the problem as follows. The joint density (or probability mass function) of a p-dimensional random vector X is characterized by parameters ϑ that depend on the edges $(s,t)\in\mathscr{E}$, for all $s,t\in\mathscr{V}$, similar to a representation of the Ising model where it is assumed that the interactions between random variables X_i are of first and second order (Wainwright and Jordan, 2008). The density of a particular node x_s conditioned upon all remaining nodes, can then be determined, based on their modelling approach as

$$f(x_s|x_{\mathscr{V}\setminus s})=\exp\{\vartheta_s x_s+\sum_{t\in\mathscr{N}_s}\vartheta_{st}x_s x_t-b(\vartheta,x_{\mathscr{V}\setminus s})+c(x_s)\},$$

where $b(\vartheta,x_{\mathscr{V}\setminus s})$ is a log-normalizing constant, $c(x_s)$ is a 'base measure' and $\mathscr{N}_s$ is the neighborhood of node s, namely the set of nodes that are directly connected to node s.

Given independent and identically distributed samples and the above conditional densities, Yang et al. (2012) then proceed by minimizing an ℓ_1-regularized condi-

tional log likelihood (see also Section 4.3) at each of the nodes, estimating sets of neighbors for each node. The merit of such an approach is that under general 'ℓ_1' regularity conditions the estimated neighbors correspond with high probability to the ones in the underlying, unknown graph, thus making the effort worthwhile and at the same time justifying why a simple nodewise approach is a sensible thing to do.

A second important interest lies in knowing if for non-Gaussian graphs, the missing edges in $\mathscr{G}$ can be translated into a '0' entry in the inverse of a covariance matrix, mimicking the behavior encountered for Gaussian graphical models. This is the topic of Loh and Wainwright (2012). Unfortunately this property of having 0's on position (s,t) and (t,s) in a general inverse of a covariance matrix if an edge is missing between nodes s and t does not hold for general graphs. Corollary 2 in their paper asserts that the inverse covariance matrix is graph structured only for graphs with singleton separator sets. Outside this condition, one can still observe 0's in a inverse covariance matrix constructed not on the original nodes but on an 'augmented' set of nodes where one includes also higher order interactions between the nodes in the set

$$S(s;d) := \{U \subseteq \mathscr{V} \setminus s, |U| = d\}$$

where d denotes an upper bound on the degree of node s (i.e. the number of edges connecting node s to any other node in the graph). As such, Corollary 3 in Loh and Wainwright (2012) asserts that the inverse of the augmented covariance matrix contains 0's on positions (s,t) for all nodes $t \notin \mathscr{N}_s$. It is in a sense a weaker result than desired, but nonetheless quite useful in understanding which conditional independencies can be read from the graph and how these are translated in 0 entries in a more familiar and easier to use generalized inverse covariance matrix. The main conclusion of the above line of work is that nodewise models can still enjoy good theoretical properties and that, as in the Gaussian case, a missing edge in $\mathscr{G}$ can still correspond to a 0 element in an augmented covariance matrix.

While in Yang et al. (2012) sparsity constraints were included and large graphs were considered, we start in Section 4 with unconstrained estimation for small to moderately sized graphs.

Extending upon the nodewise linear regression models in (1) and (2), we will include an edge in the graph between nodes i and j when using the focused information criterion, see Section 4, results in including variable X_j in the generalized linear model using X_i as a response variable with a non-zero coefficient β_{ij},

$$\mathrm{g}\{\mathrm{E}(X_i|\{X_j : j \in \mathscr{V} \setminus i\})\} = \beta_{i0} + \sum_{l \in \mathscr{V} \setminus i} \beta_{il} X_l,$$

and vice versa, when β_{ji} is nonzero in the generalized linear model when X_j is the response variable. In the case that X_i is a binary random variable, logistic regression models may be used to model the log-odds

$$\log\left\{\frac{P(X_i = 1|\{X_j : j \in \mathscr{V} \setminus i\})}{1 - P(X_i = 1|\{X_j : j \in \mathscr{V} \setminus i\})}\right\} \equiv \log\left\{\frac{\pi_i}{1-\pi_i}\right\} = \beta_{i0} + \sum_{l \in \mathscr{V} \setminus i} \beta_{ik} X_l.$$

The 'response' node is referred to as the 'child' and the 'covariate' nodes with non-zero coefficient are commonly called the 'parents' of that node.

4 The focused information criterion for graphs

As a definition of a focus, the current theoretical derivation allows for any function μ of the nodewise model parameters β that is differentiable with respect to these parameters, at least in a neighborhood of the true but unknown parameter values β_0. We will here define the focus nodewise such that we can readily apply existing search algorithms for nodewise variable selection. The mean squared errors of nodewise focus estimators are summed to yield the graph-wise mean squared error (Pircalabelu et al., 2012).

4.1 Model notation and local misspecification

Consider a sample of n observations of the p-variate vector $X_k = (X_{k1}, \ldots, X_{kp})$, with $k = 1, \ldots, n$. For each node $j = 1, \ldots, p$ and for each observation $k = 1, \ldots, n$, we have that

$$X_{kj} | \{X_{ki} : i \in \mathscr{V} \setminus j\}$$

follows a generalized linear model as in (3).

We define the vector θ_j to contain the parameters that should be estimated in all models for this node and that are always included. One example is the scale parameter ϕ when not already specified by the particular exponential family distribution. The vector θ_j might also include the coefficient corresponding to parent nodes that are forced to be in the graph, often based on domain knowledge or on theoretical grounds. Note that θ_j may be empty (absent).

We further define for each node $j \in \mathscr{V}$ the vector γ_j of length $p-1$ with ith element, $i \in \mathscr{V}$, equal to

$$\gamma_{ji} = \begin{cases} \beta_{ji} \text{ if } X_i \text{ is a parent of } X_j \\ 0 \quad \text{otherwise.} \end{cases}$$

Note that the vector γ_j does not have any overlap in parameters with θ_j, that is, model parameters are either included in γ_j or in θ_j. Thus for each node $j \in \mathscr{V}$ the vector of unknown parameters $\beta_j = (\theta_j, \gamma_j)$.

This notation assumes that for every node a full model is fit, with all other nodes as parents. This results in a full graph, where all nodes are connected to all other nodes. It is the task of a model selection method such as the FIC that we will use, to properly select the parents of each node, and as such, to reduce the fully connected graph to a simpler graph.

For this purpose we introduce notation for submodels. For each node $j \in \mathscr{V}$, when using a submodel $S \subset \mathscr{V} \setminus j$, we denote by γ_S the subvector of γ formed by the components $\{\gamma_{ji} : i \in S\}$. In the submodel defined by S, other components γ_{jk} with $k \notin S$ are taken to be zero. Such a selection of components may algebraically be defined through multiplication with projection matrices selecting the wanted components, see Claeskens and Hjort (2008b, sec. 6.1).

The nodewise focus μ_j that we wish to estimate with low mean squared error can be written as $\mu_j(\theta_j, \gamma_j; x)$. One example is a nodewise expectation $\mu_j(\theta_j, \gamma_j; x) = x^t \gamma_j$ for a user-specified vector x. We will estimate μ_j in a submodel S by $\hat{\mu}_{j,S} = \mu_j(\hat{\theta}_{j,S}, \hat{\gamma}_{j,S}; x)$ using maximum likelihood estimators in the submodel. Note that no selection of components takes place for θ_j, though its estimated value might in general depend on which components of γ_j are included in S.

In order to estimate the mean squared error of $\hat{\mu}_S$, we employ a *local misspecification framework* where the true parameter vector has the form $(\theta_{j,0}, \gamma_{j,0} + \delta_j/\sqrt{n})$, for some unknown vector δ. This construction will result in squared biases of estimators that are of the same order as variances, thus resulting in mean squared error values that are not driven by bias or variance only, as the sample size grows. Working under a fixed true model (not depending on the sample size) would lead to suggest to always use the full model since in that case the bias would dominate, see Claeskens and Hjort (2008b, sec. 5.2).

4.2 FIC for small to moderate graphs

The strategy for estimation of the mean squared error of $\hat{\mu}_{j,S}$ in each considered model S is as follows. By taking the mean and variance from the asymptotic distribution of the estimator $\hat{\mu}_{j,S}$, the mean squared error is easily formed. This expression is estimated in a next step to form the focused information criterion. For the asymptotic distribution of the estimators $\hat{\mu}_{j,S}$ under local misspecification in the specific case of generalized linear models, see Claeskens and Hjort (2008a).

Let us consider the general situation where there is an unknown scale parameter ϕ in the exponential family distribution and where some of the parents are protected from variable selection. For node $j \in \mathscr{V}$, denote the 'protected' parents, those that are forced to be present in the graph, by U_j and those that are subject to model selection by Z_j; remark that nodes are either protected or unprotected, not both, hence U_j and Z_j do not contain common components. Likewise, we write $x = (u, z)$.

Define $J_{n,\phi} = -n^{-1} \sum_{k=1}^{n} E[\partial^2 \log f(X_{kj}; \vartheta_k, \phi)/\partial \phi^2]$ using the exponential family density function as in (3). Then, the information matrix corresponding to the full model for the jth node is given by

$$J_n = \begin{pmatrix} J_{n,\phi} & 0 & 0 \\ 0 & n^{-1}a(\phi)^{-1}U^tVU & n^{-1}a(\phi)^{-1}U^tVZ \\ 0 & n^{-1}a(\phi)^{-1}Z^tVU & n^{-1}a(\phi)^{-1}Z^tVZ \end{pmatrix},$$

for which we assume that a limit J exists for n tending to infinity; this condition could also have been phrased in terms of conditions on the design matrices U and Z. We assume that J_n and J are invertible. The matrix V that is used in J_n is a diagonal matrix $\mathrm{diag}\{v_1,\ldots,v_n\}$ with $v_k = [b''(\vartheta_k)\{g'(\xi_k)\}^2]^{-1}$ and $\xi_k = E[X_{jk}|U_{jk},Z_{jk}] = b'(\vartheta_k)$. The vector θ_j consists of ϕ and of the coefficients ν_j belonging to the protected variables U_j. Thus $\theta_j = (\phi,\nu_j)$. Denote the length of θ_j by p_θ and the number of elements in S by $|S|$.

Standard maximum likelihood methods yield that as n tends to infinity,

$$\begin{pmatrix}\sqrt{n}(\hat\theta_{j,S}-\theta_{j,0})\\ \sqrt{n}\hat\gamma_{j,S}\end{pmatrix} \xrightarrow{d} N_{p_\theta+|S|}\left(\begin{pmatrix}0\\ \delta\end{pmatrix}, J_S^{-1}\right).$$

We denoted by J_S^{-1} the inverse of the $(p_\theta+|S|)\times(p_\theta+|S|)$ matrix J_S that is formed by selecting from J those rows and columns indexed by S.

Since we are interested in the asymptotic distribution for the focus estimator at the jth node $\mu_j(\hat\theta_{j,S},\hat\gamma_{j,S};x)$, we use the differentiability of μ_j with respect to the parameters (θ,γ) to first define

$$\begin{aligned}\omega &= Z^tVU(U^tVU)^{-1}\frac{\partial\mu_j}{\partial\nu_j} - \frac{\partial\mu_j}{\partial\gamma_j},\\ \tau_0^2 &= \frac{1}{J_\phi}\left(\frac{\partial\mu_j}{\partial\phi}\right)^2 + na(\phi)\left(\frac{\partial\mu_j}{\partial\nu_j}\right)^t(U^tVU)^{-1}\left(\frac{\partial\mu_j}{\partial\nu_j}\right),\end{aligned}$$

where all partial derivatives are evaluated at $(\theta_0,0)$. Then, see Claeskens and Hjort (2008b, Chapter 6) and Claeskens and Hjort (2008a), as n tends to infinity,

$$\sqrt{n}(\hat\mu_{jS}-\hat\mu_{j\mathrm{true}}) \xrightarrow{d} \Lambda_S,$$

where $\mathrm{E}(\Lambda_S) = \omega^t(I_{p-1}-G_S)\delta$ and $\mathrm{Var}(\Lambda_S) = \tau_0^2+\omega^tQ_S^0\omega$ with Q the limit of $Q_n = a(\phi)n\{S^tV(I_n-U(U^tVU)^{-1}U^tV)Z\}^{-1}$, I_n a square identity matrix with n rows and G_S the limit of $G_{n,S} = Q_{n,S}^0Q_n^{-1}$. The matrix $Q_{n,S}^0$ is defined as follows. Take from Q_n^{-1} the submatrix consisting of those rows and columns indexed by S. We invert the obtained matrix and place its matrix elements in a $(p-1)\times(p-1)$ matrix in the rows and columns indexed by S, and set the other matrix elements equal to zero. In words, Q^{-1} is premultiplied with part of its inverse such that $G_{n,S}$ is a zero matrix when S is the empty set and $G_{n,S}$ is the identity matrix for the full model when $S = \mathcal{V}\setminus j$. Since $G_{n,S}$, $Q_{n,S}$ and $Q_{n,S}^0$ are all defined via submatrices of J_n, the existence of a limit matrix for $n\to\infty$ is guaranteed via the existence of the limit matrix J and of its inverse J^{-1}.

We now obtain the mean squared error for $\mu_j(\hat\theta_{j,S},\hat\gamma_{j,S})$ by adding its variance and its squared bias as

$$\mathrm{MSE}(\hat\mu_{jS}) = \tau_0^2+\omega^tQ_S^0\omega+\omega^t(I_{p-1}-G_S)\delta\delta^t(I_{p-1}-G_S)^t\omega, \tag{4}$$

where Q_S^0 is the limit of $Q_{n,S}^0$ and I_{p-1} represents a square identity matrix with $p-1$ rows. The best choice of parents to use in the nodewise regression model is that set S for which $\text{MSE}(\hat{\mu}_{jS})$ is as small as possible. Since this expression contains several unknown quantities, we insert estimators for unknowns, indicated by a 'hat' notation, where for example $\hat{Q}$, $\hat{Q}_S^0$ and $\hat{G}_S$ represent the empirical estimates of the corresponding matrices, resulting in an expression for the focused information criterion, FIC.

In particular, we estimate $\delta\delta^t$ unbiasedly by $\hat{\gamma}_{j,w}\hat{\gamma}_{j,w}^t - \hat{Q}$ where $\hat{\gamma}_{j,w}$ is the estimator of γ_j in the wide, or full, model using $S = \mathscr{V} \setminus j$, and an empirical information is used with parameters estimated at the full model. This results in defining the focused information criterion for node $j \in \mathscr{V}$ using subset S as parents:

$$\text{FIC}(S,\mu_j) = \hat{\tau}_0^2 + 2\hat{\omega}^t\hat{Q}_S^0\hat{\omega} + n\hat{\omega}^t(I_{p-1} - \hat{G}_S)\hat{\gamma}_{j,w}\hat{\gamma}_{j,w}^t(I_{p-1} - \hat{G}_S)^t\hat{\omega} - \hat{\omega}^t\hat{Q}\hat{\omega}. \quad (5)$$

Note that since the first and the last term do not depend on the particular submodel S, these terms may be omitted when nodewise ranking the values of $\text{FIC}(S,\mu_j)$ for different sets S. Further, note that in these nodewise regression models, also the matrix Q_n, and as a consequence also $\omega, Q_{n,S}^0$ and $G_{n,S}$ are nodewise defined.

The value of the FIC for the complete graph is defined by Pircalabelu et al. (2012) as the nodewise summation of the FIC values for each node given in (5),

$$\text{FIC}(\mathscr{S};\mathscr{G}) = \sum_{j=1}^{p} \text{FIC}(S_j,\mu_j), \quad (6)$$

where $\mathscr{S} = \{(S_1,\ldots,S_p) : S_1 \subseteq \{\mathscr{V} \setminus 1\};\ldots;S_p \subseteq \{\mathscr{V} \setminus p\}\}$, and each S_j corresponds to the selected nodes that minimize the FIC score of the estimated focus at node j.

The best graph in estimated MSE sense according to the focused information criterion is given by that selection of nodewise parents S_j $(j = 1,\ldots,p)$ for which the combined FIC value $\text{FIC}(S;\mathscr{G})$ is the smallest over all considered sets. In the case it happens that two choices of $\mathscr{S}$ would give identical FIC values, other aspects of modeling, e.g. parsimony considerations, might help decide the final selection, in the same way as is done for model selection via other information criteria. Model averaging might also be an option when prediction is the objective.

4.3 FIC for large graphs

While the FIC in (6) relies on maximum likelihood estimation, this no longer is feasible when many nodes are involved. For situations with many unknown parameters (including the situations where there are more unknown parameters than observed cases), penalized estimation methods are appropriate.

In such case the estimators are maximizers of the penalized objective function

$$Q(\theta,\gamma) = \frac{1}{n}\sum_{k=1}^{n} \log f(y_k|w_k,z_k,\theta,\gamma) - \frac{1}{n}\sum_{j=1}^{p-1} \psi_\lambda(|\gamma_j - \gamma_{j0}|), \tag{7}$$

with respect to θ and γ for a given penalty function ψ that is twice differentiable in 0 and that depends on the penalty constant λ. This $\lambda \geq 0$ is a user-determined value, which may be obtained in a data-driven fashion, and γ_{j0} is the value of the coefficient γ_j in the narrow model. The effect of the penalty is that the estimators are shrunk towards zero. Typical choices are ℓ_2 (sum of squares), ℓ_1 (sum of absolute values) or ℓ_0 (hard thresholding) penalties.

By adding a penalty to the estimation Meinshausen and Bühlmann (2006) propose using a series of nodewise Lasso regression models, using an ℓ_1 penalty, to estimate large graphical models. See also Wainwright et al. (2007) and Schmidt et al. (2007) among many others. Neighborhoods of different nodes can be connected in an undirected graphical structure by means of an 'AND' rule, or an 'OR' rule, in the same way as for unpenalized nodewise regression models,

$$\hat{\mathscr{E}}_\lambda^{\text{AND}} = \{(i,j) : i \in \widehat{\mathscr{N}_j}(\lambda) \text{ AND } j \in \widehat{\mathscr{N}_i}(\lambda)\},$$

$$\hat{\mathscr{E}}_\lambda^{\text{OR}} = \{(i,j) : i \in \widehat{\mathscr{N}_j}(\lambda) \text{ OR } j \in \widehat{\mathscr{N}_i}(\lambda)\}.$$

For non-differentiable penalty functions, such as the ℓ_1 or ℓ_0 penalties, which are not differentiable at zero, Fan and Li (2001) suggest a local quadratic approximation. This had lead Pircalabelu et al. (2013) to use the following approximations to $\psi_\lambda(|\gamma_j - \gamma_{j0}|)$, $\psi'_\lambda(|\gamma_j - \gamma_{j0}|)$ and $\psi''_\lambda(|\gamma_j - \gamma_{j0}|)$, where γ_{japx} is a value close to $|\gamma_j - \gamma_{j0}|$,

$$\begin{aligned}
\psi_\lambda(|\gamma_j - \gamma_{j0}|) &\approx \psi_\lambda(\gamma_{j\,\text{apx}}) + \tfrac{1}{2}\frac{\psi'_\lambda(|\gamma_{j\text{apx}}|)}{|\gamma_{j\text{apx}}|}\left[(\gamma_j - \gamma_{j0})^2 - \gamma^2_{j\text{apx}}\right];\\
\psi'_\lambda(|\gamma_j - \gamma_{j0}|) &\approx \frac{\psi'_\lambda(|\gamma_{j\text{apx}}|)}{|\gamma_{j\text{apx}}|}(\gamma_j - \gamma_{j0});\\
\psi''_\lambda(|\gamma_j - \gamma_{j0}|) &\approx \frac{\psi'_\lambda(|\gamma_{j\text{apx}}|)}{|\gamma_{j\text{apx}}|}
\end{aligned}.$$

The above quadratic approximations have been used on the one hand to 'ease' the optimization problem by making use of a relatively fast iterative procedure in order to obtain estimated coefficients. On the other hand, more importantly, they have been introduced to satisfy the existence of a second derivative at zero, needed in (8), which is not generally satisfied by most penalty functions. Working with non-differentiable expressions might lead to an alternative approach to obtain the MSE that avoids such approximations, however, this is not addressed here.

In the practical computations, the value $\gamma_{j\text{apx}}$ is arbitrarily at the start and is updated in an iterative Newton-Raphson scheme.

Often used examples of penalty function that can be used in (7) with these approximations include

- lasso (Tibshirani, 1996) with $\psi^l_\lambda(|\gamma_j - \gamma_{j0}|) = \lambda|\gamma_j - \gamma_{j0}|$;
- bridge (Frank and Friedman, 1993) with $\psi^b_\lambda(|\gamma_j - \gamma_{j0}|) = \lambda|\gamma_j - \gamma_{j0}|^\alpha;\ \alpha > 0$;

- hard thresholding: $\psi_\lambda^h(|\gamma_j-\gamma_{j0}|)=\lambda^2-(|\gamma_j-\gamma_{j0}|-\lambda)^2 I(|\gamma_j-\gamma_{j0}|<\lambda)$;
- adaptive lasso (Zou, 2006) with $\psi_\lambda^{al}(|\gamma_j-\gamma_{j0}|)=\lambda w_j|\gamma_j-\gamma_{j0}|$ with w_j being a set of weights corresponding to each node in the graph ;
- Smoothly clipped absolute deviation (SCAD, Fan and Li, 2001) for which the first derivative is defined as
$\psi_\lambda^{'s}(|\gamma_j-\gamma_{j0}|)=I(|\gamma_j-\gamma_{j0}|\leq\lambda)+\frac{(a\lambda-|\gamma_j-\gamma_{j0}|)_+}{(a-1)\lambda}I(|\gamma_j-\gamma_{j0}|>\lambda); a>2.$

The nodewise MSE for the estimator for μ_j in model S can for penalized estimation be written as (Pircalabelu et al., 2013)

$$\begin{aligned}\text{MSE}(\hat{\mu}_{jS}) = \tau_0^2+\omega^t Q_S^0\omega+\omega^t\{(I_{p-1}-G_S)\delta\delta^t(I_{p-1}-G_S^t)\}\omega+\\ +\omega^t\{Q_S^0 cc^t(Q_S^0)^t-2(I-G_S)\delta c^t(Q_S^0)^t\}\omega,\end{aligned} \tag{8}$$

where $c=n^{-1/2}\psi_\lambda^{''}(0)1_{p-1}$. Note that (8) reduces to (4) when there is no penalty, thus $\lambda=0$. The FIC for penalized estimation results by inserting in (8) estimators obtained in the full model for unknowns, leading to

$$\begin{aligned}\text{FIC}(S,\mu_j;\lambda) = \hat{\tau}_0^2+2\hat{\omega}^t\hat{Q}_S^0\hat{\omega}+n\hat{\omega}^t(I_{p-1}-\hat{G}_S)\hat{\gamma}_{j,w}\hat{\gamma}_{j,w}^t(I_{p-1}-\hat{G}_S)^t\hat{\omega}-\hat{\omega}^t\hat{Q}\hat{\omega}\\ +\hat{\omega}^t\{\hat{Q}_S^0 cc^t(\hat{Q}_S^0)^t-2(I-\hat{G}_S)n^{1/2}\hat{\gamma}_{j,w}c^t(\hat{Q}_S^0)^t\}\hat{\omega}.\end{aligned} \tag{9}$$

Note that the value of the FIC depends on the choice of λ (which is contained in c). Since the above procedure is applied to each node, the modeling strategy allows thus different amounts of penalization at each node. In practice, for each node a value from a grid of λ values is chosen based on a three-fold cross-validation procedure on the deviance of the GLM.

5 Computational aspects

While for small graphs only containing a small number of nodes it might be possible to investigate an all subsets search for each node, this is not feasible for moderate to large sized graphs.

For large graphs, at each node a penalized GLM based on all the other nodes is fitted from which one obtains immediately the penalized maximum likelihood estimator $(\hat{\theta},\hat{\gamma})$ in the full model as well as the empirical Fisher information matrix J_n and the weights for the 'working-variables' once the Newton-Raphson algorithm converges. By allowing for a quadratic approximation of the penalty, the problem which was originally a convex problem, now enjoys first and second order differentiability properties as well, making the optimization based on Newton-type methods easy to implement.

Once all the necessary quantities have been estimated from the full model, we start building collections of models S in an incremental fashion. We start first from an intercept-only model (for which the cardinality is 1) and compute its FIC score. In a second step, all models that include one extra neighbor (having thus cardinality

2) are compared to the benchmark model, namely the intercept model. The model with the lowest FIC score then becomes the new benchmark. If none of the models provides lower FIC values than the intercept model, the procedure stops. Otherwise, all models of cardinality 3 that include the benchmark model, are compared to the benchmark model. If any of these models, improves the FIC score we retain it and then search for models with higher cardinality, otherwise the procedure stops and outputs the model with the best attained score so far. Since this is a greedy local search algorithm and since the relation between the FIC scores and cardinality is non-linear, we do not restrict to making every time the hard decision of stopping the search if the score is not improved at each step, but test also some locally non-optimal models which at the next stage due to the inclusion of other neighbors, might offer a better FIC value than if we would have stopped at the best model from the previous stage.

6 Data analysis

We here return to data examples stated in Section 2.

6.1 Dynamics of psychopathology

Due to the binary recoding of the data, seven of the 22 items in that PANAS resulted in having constant values (all zero, or all one) across the 93 days, these items have been excluded from the analysis. After this elimination we have treated each of the remaining items as a node in an undirected network, where the goal was to discover the edges that provide the lowest MSE for the $\text{logit}(\pi_i)$ where π_i is the probability that item (or node) i indicates a tendency towards negative feelings.

The datasets for the two subjects were treated separately, in two distinct applications of the same FIC procedures. The observed sequence of emotions provides information on how a patient (or control) is doing. Therefore, we specified the focus point as being the observed sequence of emotions at each day. Afterwards, for each of the specified focuses we estimate a network and based on that network we estimated the basic reproduction number ρ, see Section 2.1. This resulted (due to missing observations) in having specified 90 different focuses and so 90 different networks (for each network a value of ρ is estimated) for the patient and 88 different focuses and networks for the control subject.

The questions for which we want to find an answer can be formulated as follows: having the entire dataset of observations for the patient (likewise for the control), and assuming that tomorrow we observe a sequence of emotions that corresponds to what we have observed at time $t \in \{1,\ldots,90\}$, what is the topology of the network which would generate a low MSE of the focus at each of the nodes? For example, Figure 1 presents four estimated networks corresponding to the sequence of emo-

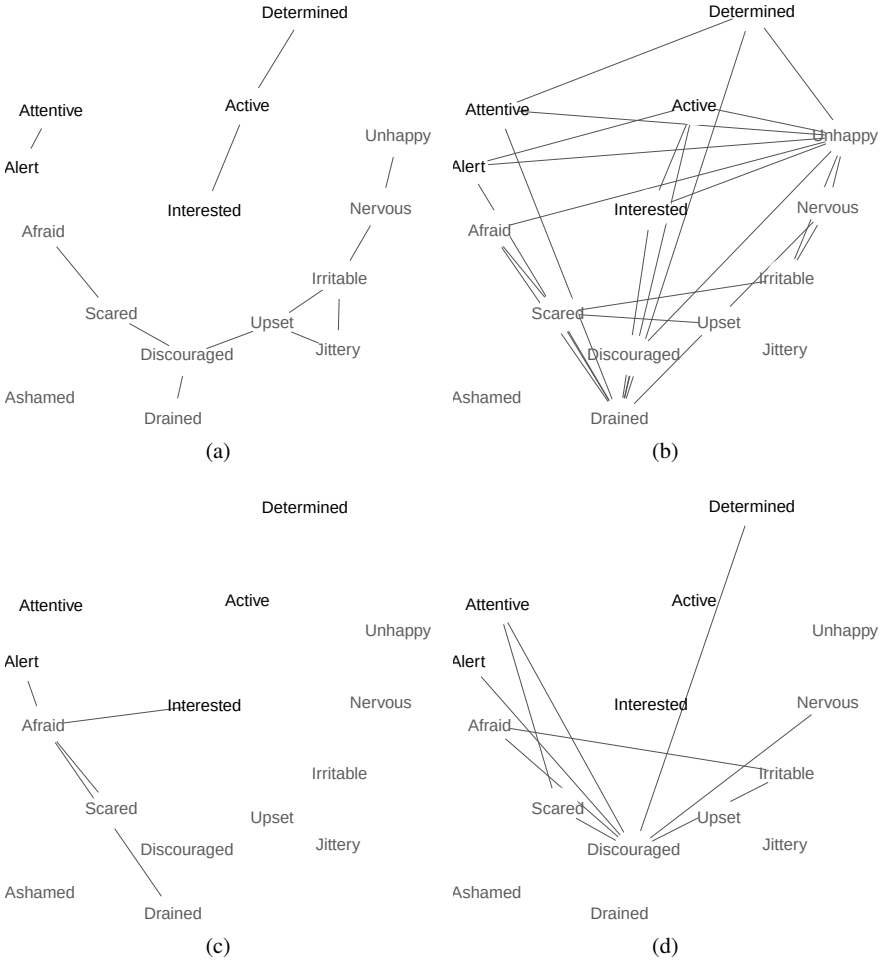


Fig. 1: PANAS data. Visual representation of the graphical structure estimated using a local quadratic approximation to an ℓ_1 penalty when the focus point is the sequence of emotions observed at time points 1 (panels a,c) and 70 (panels b,d) for the patient (panels a,b) and control (panels c,d). The corresponding estimated BRP rates are equal to 5.81 and 1.31 for the patient and 0.99 and 0.78 for the control. The black (resp. gray) colors reflect the positive (resp. negative) affect aspect of the node.

tions that were observed for both the patient and the control, at time points 1 and 70. It is apparent that for the first time point in the estimated graphs for the patient there is a higher tendency to separate the negative affects from the positive ones, whereas in the graphs estimated at the second time point there is a tendency to have a higher density of edges and to also positive and negative affects get linked with each other much more often blurring in a sense the separation between the two categories of

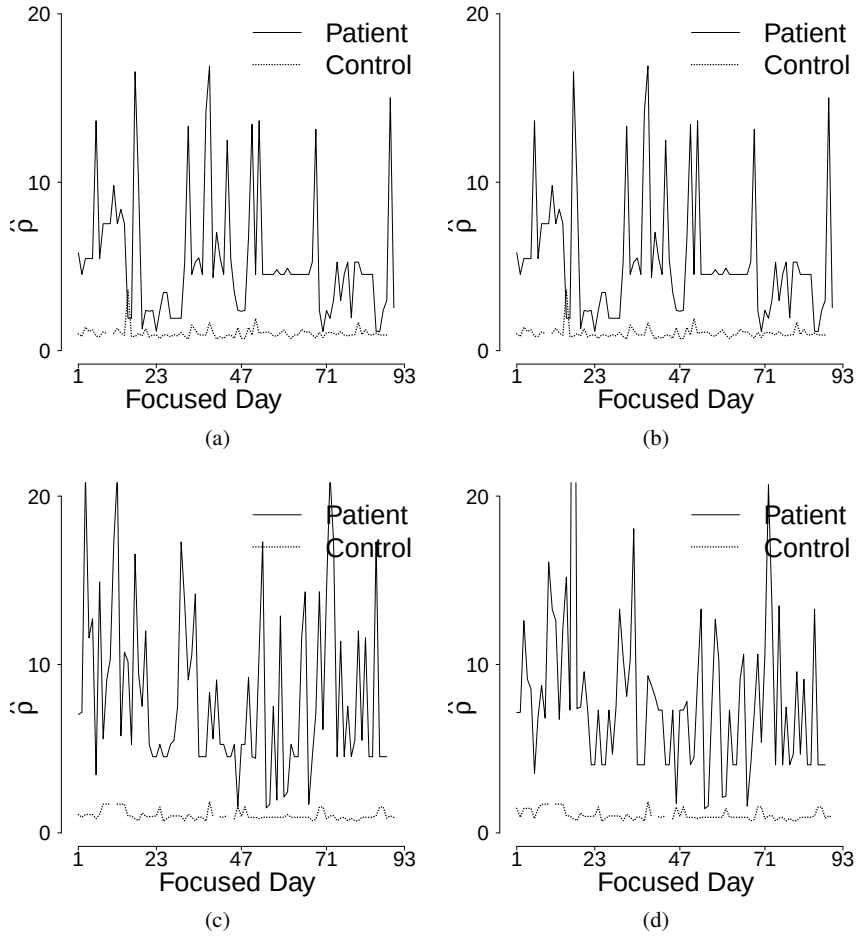


Fig. 2: PANAS data. Plotted is the $\hat{\rho}$ for each of the two subjects, when different observed sequence of emotions (corresponding to a day on the x-axis) are chosen as focus points. In the upper row, the focus points come from the same subject as the training data, whereas in the second row the focus points come from the emotion pattern displayed by the other subject. The local quadratic approximation to an ℓ_1 penalty (left column) and to a SCAD penalty (right column) have been used for estimating the corresponding networks for the two subjects.

feelings. As expected the network topology plays a crucial role, as for instance the four estimated ρ's based on these networks are quite different, especially for the patient which for these two focuses exhibits higher BRPs.

Since one can estimate thus a multitude of networks, each pertaining to the sequence of emotions observed on a particular day, one might also be interested in how 'stable' the patient tends to have a high BRP. Is this phenomenon stable across

sequences (one for each day) of affects observed at each time point or was the above conclusion largely due to the effect of the particular focuses? To answer this question we have plotted the estimated BRPs for both subjects as a function of time. The results are presented in the upper row of Figure 2 and it is apparent that the levels of the observed trends are almost always larger for the patient than for the control across many such observed sequences. Quite interestingly, this analysis seems to support the conclusions of the original authors concerning the fact that the patient exhibits higher BRP rates than the control, though coming from a conceptually different stand point with respect to estimating an unknown hidden undirected network.

A further investigation concentrates on 'confusing' the FIC procedure in the following sense. Up to now both the data used for the estimation as well as the focus would come from the same subject in a sense making the problem somewhat easier. As such we were interested in the discriminatory power of the procedure when the data came from the patient, but the focus point came from the control. This would correspond to the situation where based on the behavior seen so far, if for a brief moment the patient would exhibit normal behavior (situation summarized by the focus point), can he still be categorized as being patient based on the estimated ρ? Or vice versa, if based on what was observed so far, if a healthy subject exhibits for a moment a sequence of emotions similar to what the patient exhibited, do we still estimate networks for which ρ is relatively large? To answer this question we have proceeded as in the above application, but with the major difference that now the focuses are coming from what was observed for the other subject. The bottom row of Figure 2 illustrates the findings and supports the conclusion that even though the FIC procedure estimated graphs that exhibited generally higher BRP ratios for the patient than for the control, it is still able to discriminate between the two subjects based on the proposed ratio, even when the focuses are probably not in line with the data used for estimation.

6.2 U.S. voting behavior

Since the vote is coded with a binary value, we fit at each node (i.e. Senator) a penalized logistic regression model with the vectors of votes for all remaining senators used as predictors. Based on this full model we construct the estimate $\hat{\omega}_w$ using the estimated vector of regression coefficients $\hat{\beta}$, corresponding to the influence of each 'covariate' or 'parent' node on the probability of a 'Yes' vote for the dependent node. The intercept of the model, β_{i0}, acts as the protected parameter θ.

In order to fix notation, let $X_{ki} \sim \text{Bernoulli}(\pi_i)$ with

$$\log\left\{\frac{\pi_i}{1-\pi_i}\right\} = \beta_{i0} + \sum_{l \in \mathscr{V}\setminus i} \beta_{il} X_{kl} - \frac{1}{n} \sum_{l \in \mathscr{V}\setminus i} \psi_\lambda(|\beta_{il}|)$$

where X_{ki} denotes the result of a vote of Senator i on bill k and X_{kl}, for $l \in \mathscr{V} \setminus i$, represent the voting results for the remaining senators on the same bill. A constrain-

ing penalty function is placed on the vector of unknown β parameters. In the narrow model all coefficients corresponding to the unprotected nodes are set equal to zero, since in the narrow model none of them is included.

In a subsequent step we proceed as in Section 4.3. At the dependent node we select from the set of potential neighbors, the ones which minimize the MSE($\hat{\mu}_S$). Once the set of neighbors is selected for a particular node, we move to another node and proceed in the same fashion by estimating its set of neighbors. We perform the same procedure at each of the $p = 100$ nodes in the graph, obtain p sets of neighbors and afterwards combine all the information by drawing an edge between two nodes i and j if i belongs to the set of neighbors of j or vice versa. Notationally, this amounts to $(i, j) \in \mathscr{E}$ if $i \in \widehat{\mathscr{N}_j}$ OR $j \in \widehat{\mathscr{N}_i}$.

Figure 3 illustrates a few interesting patterns. First of all, it seems that the 'party vote' had a major role to play as most of the edges in the graph link two senators that belong to the same party. Second, within the Democratic party, the graph suggests that senators opposing the amendment are more likely to get linked to other democrats opposing the amendment than to the democrats in favor of the amendment. Lastly, the graph suggests also that there are some between-party edges, although they appear less frequently than the within-party edges.

Since at each node i, neighbors are selected on the basis that the model provides the lowest estimated MSE for logit(π_i) where π_i is the probability that bill i receives a favorable vote, one might be interested in the performance of such a classifier for the focus for which it was constructed. In this case this corresponds to predicting for each senator his vote on the bill. Based on the graph presented in Figure 3 we estimate the correct vote for 78% (or 84 % for SCAD) of the senators, whereas predicting based on average vote for all other bills (not incorporating any knowledge about the relations between senators) resulted in a correct prediction in only 46% of the cases. These predictions are slightly optimistic since they are within sample, as the information which we are predicting is also used for constructing the graph.

6.3 Hunting spider species

Since the number of captured spiders is observed per location, we estimate an interactions network where connected nodes indicate that the two species are co-occurring. At each node a Poisson model is fitted where $X_{ki} \sim \text{Poisson}(\xi_i)$ with

$$\log(\xi_i) = \beta_{i0} + \sum_{l \in \mathscr{V} \setminus i} \beta_{il} X_{kl} - \frac{1}{n} \sum_{l \in \mathscr{V} \setminus i} \psi_\lambda(|\beta_{il}|)$$

where X_{ki} denotes the number of spiders at location i coming from species k and X_{kl}, for $l \in \mathscr{V} \setminus i$, represents the number of spiders at the remaining locations coming from the same species. A constraining penalty function is placed again on the vector of unknown β parameters, and in the narrow model all coefficients corresponding to the unprotected nodes are set equal to zero.

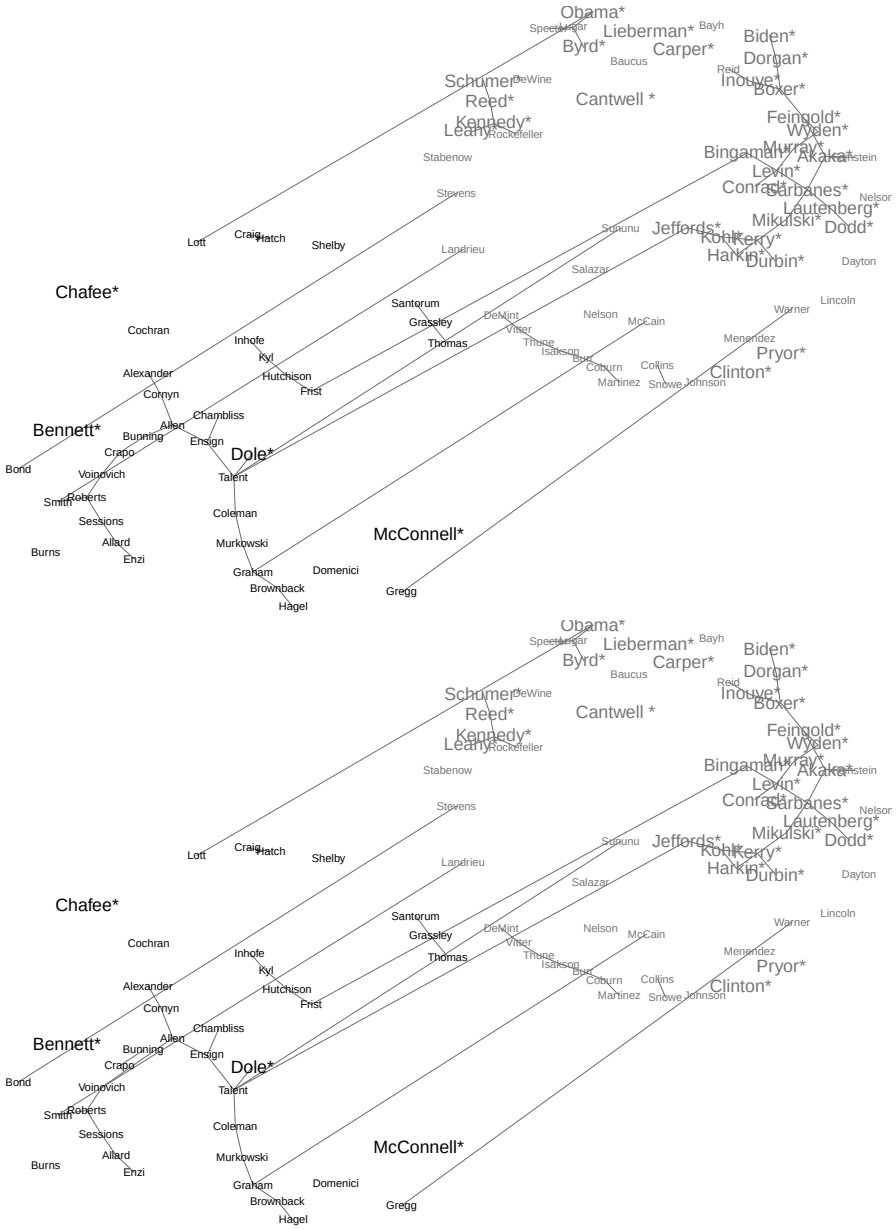
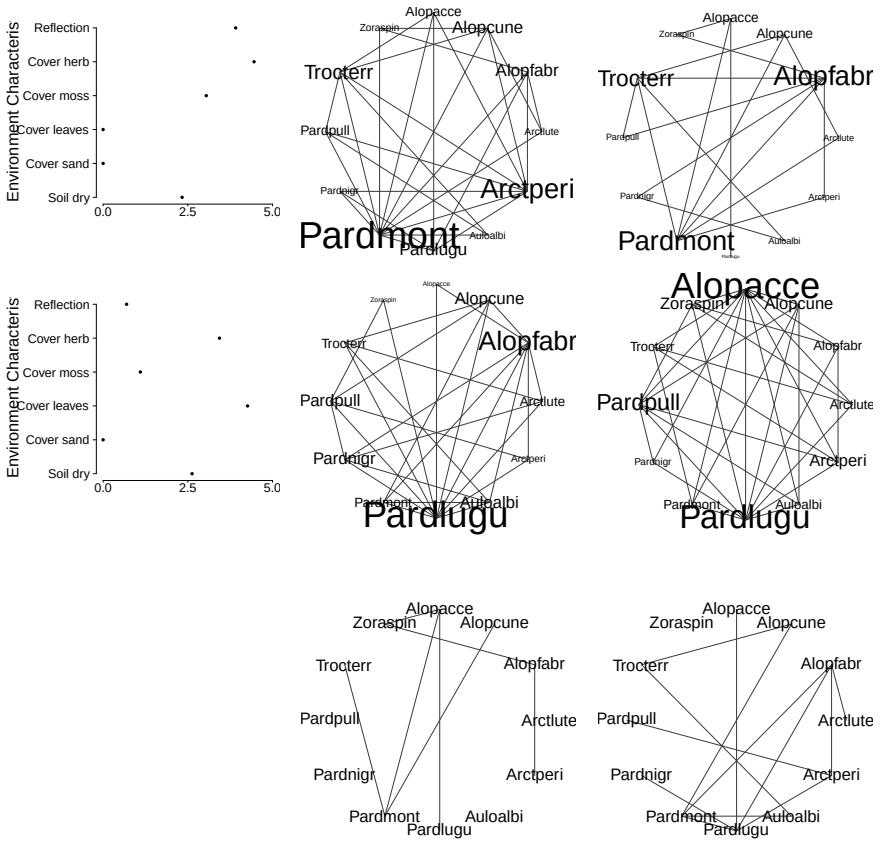


Fig. 3: Senate vote data. Visual representation of the graphical structure estimated using a local quadratic approximation to an ℓ_1 penalty (top) and to a SCAD penalty (bottom). In each figure, black nodes denote the Republican senators (lower left quadrant) and the gray nodes denote the Democrat senators (upper right quadrant). Within each party, the nodes accompanied by a $\star$ symbol denote senators that have opposed the amendment.



(a) Focus (b) FIC LQA Bridge (c) FIC ℓ_2

Fig. 4: Spider data. FIC estimated graphs (first two lines) based on two focuses (the corresponding environmental characteristics are shown in the left column). The ℓ_2, LQA Bridge denote the ψ function used in the estimation. Larger labels correspond to highly connected nodes and the bottom line presents the common edges across estimated graphs per focus.

Figure 4 shows the estimated undirected graphs for two focuses when (i) a quadratic approximation of the Bridge penalty and (ii) an ℓ_2 penalty is used. The immediate conclusion is that the characteristics of the environment have an influence on the structure of the estimated networks as does the type of penalty that is being used. For the first focus both graphs suggest that the Pardmont species tends to co-occur most often with the other species, while for the second focus the Pardlugu species is the highly connected species. The ℓ_2 graph identifies for the second focus also the Alopacce species as being highly connected to other species as well. Quite interesting is the fact that regardless of the environment conditions for the two cases

studied the pairs Pardlugu-Alopacce, Pardmont-Allopcune and Alopfabr-Arctperi are present in all graphs and their abundance seems to be related.

7 Discussion

In this chapter we presented an extension of our method to construct graphs from the focused information criterion to generalized linear models. The three main advantages of using the FIC to construct graphs are: (i) a focus of interest can be defined incorporating prior knowledge of the system under investigation, (ii) the mean squared error of the focus is minimized which balances squared bias and variance of the estimator and increases generalizability, and (iii) the framework of local misspecification is used relaxing the assumption of having the correct model.

We showed that the combination of the GLM and FIC leads to an easily interpretable Fisher information matrix, separating the two types of parameters, ones that are always included and ones that are to be determined. This in turn was seen to lead to an estimate of the mean squared error that is used to determine the FIC.

The three examples shown in this chapter indicate the richness of the method. In the first example data from a bipolar patient and a control were contrasted suggesting different patterns of predictions for whether symptoms of bipolar disorder would remain or not. Especially interesting was the fact that using a sequence of emotional items (knowledge of the system), the basic reproduction number ρ, resulting from the estimated graph, was seen to vary strongly in the patient but not the control. And even using an emotional item sequence of the control in the bipolar patient resulted in a largely varying pattern of values of ρ. These results show that a network of emotional states which influence each other can be obtained, from which behavior of symptoms can be predicted.

The second example using data from the voting behavior of U.S. senators showed that for the 'Flag Desecration' amendment predicting voting behavior using estimated relations between senators may result in higher accuracy than using previous voting behavior of senators, while in the same time discovering that intra-party cooperation is dominant (the voting pattern of a senator can best be described by patterns of colleagues from the same party), the cluster of opposing democrats stands out in this respect, but also that cross-party cooperations is not negligible.

The third examples uses Poisson distributions to model counts of different species of spiders at different locations.

Because of the extensions to the more general exponential family of distributions enlarge the range of applicability of this procedure to the more realistic situations where one has at disposal binary or count data, the estimation of connections is not limited just to Gaussian data.

In conclusion, there are many possibilities of using the focused information criterion to obtain meaningful graphs of many kinds of systems, as shown by the examples presented here which showed that the presented FIC procedure can be useful for estimating graph structures by taking the researcher's objectives more closely

into account and outputting a model that comes closer to his goals. Since we can easily incorporate knowledge of a system through the focus, the method is flexible and useful.

Acknowledgements

The authors wish to thank Prof. J.-H. Kamphuis for the PANAS data. The authors acknowledge the support of the Fund for Scientific Research Flanders, KU Leuven grant GOA/12/14 and of the IAP Research Network P7/06 of the Belgian Science Policy.

References

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516.

Borsboom, D., Cramer, A. O. J., Schmittmann, V. D., Epskamp, S., and Waldorp, L. J. (2011). The small world of psychopathology. *PLoS ONE*, 6(11):e27407.

Claeskens, G. and Hjort, N. (2003). The focused information criterion. *Journal of the American Statistical Association*, 98:900–916. With discussion and a rejoinder by the authors.

Claeskens, G. and Hjort, N. (2008a). Minimising average risk in regression models. *Econometric Theory*, 24:493–527.

Claeskens, G. and Hjort, N. (2008b). *Model selection and model averaging*. Cambridge University Press, Cambridge.

Cox, D. R. and Wermuth, N. (1996). *Multivariate dependencies: Models, analysis and interpretation*. Chapman & Hall, London.

Dempster, A. (1972). Covariance selection. *Biometrics*, 28(1):157–175.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Fiocco, M. and van Zwet, W. (2004). Maximum likelihood estimation for the contact process. *Lecture Notes-Monograph Series*, pages 309–318.

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135.

Jalali, A., Ravikumar, P., Vasuki, V., and Sanghavi, S. (2010). On learning discrete graphical models using group-sparse regularization. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*.

Lauritzen, S. (1996). *Graphical models*. Oxford University Press.

Lee, J. and Hastie, T. (2012). Learning mixed graphical models. *ArXiv e-prints*.

Loh, P. L. and Wainwright, M. J. (2012). Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. *ArXiv e-prints*.

McCullagh, P. and Nelder, J. (1989). *Generalized linear models*. Chapman & Hall, London.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462.

Pircalabelu, E., Claeskens, G., Jahfari, S., and Waldorp, L. (2013). Focused information criterion for graphical models. large p, small n considerations. Technical report, KBI, Faculty of Economics and Business, KU Leuven.

Pircalabelu, E., Claeskens, G., and Waldorp, L. (2012). Structure learning using a focused information criterion in graphical models. Technical report, KBI, Faculty of Economics and Business, KU Leuven.

Schmidt, M., Niculescu-Mizil, A., and Murphy, K. (2007). Learning graphical model structure using ℓ_1-regularization paths. In *Proceedings of the 22nd national conference on Artificial intelligence*, volume 2, pages 1278–1283.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (SERIES B)*, 58:267–288.

van Borkulo, C. D., Kamphuis, J. H., and Waldorp, L. J. (2013). Predicting behaviour of networks of mental disorders: The contact process as a model for dynamics of psychopathology. Technical report.

van der Aart, P. J. M. and Smeenk-Enserink, N. (1975). Correlations between distributions of hunting spiders (lycosidae, ctenidae) and environmental characteristics in a dune area. *Netherlands Journal of Zoology*, 25:1–45.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305.

Wainwright, M. J., Ravikumar, P., and Lafferty, J. D. (2007). High-dimensional graphical model selection using ℓ_1-regularized logistic regression. In Schölkopf, B., Platt, J., and Hoffman, T., editors, *Advances in Neural Information Processing Systems 19*, pages 1465–1472. MIT Press.

Watson, D., Clark, L. A., and Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of Personality and Social Psychology*, 54(6):1063–1070.

Yang, E., Ravikumar, P. K., Allen, G. I., and Liu, Z. (2012). Graphical models via generalized linear models. In Bartlett, P., Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 25*, pages 1367–1375.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

FACULTY OF ECONOMICS AND BUSINESS
Naamsestraat 69 bus 3500
3000 LEUVEN, BELGIË
tel. + 32 16 32 66 12
fax + 32 16 32 67 91
info@econ.kuleuven.be
www.econ.kuleuven.be