Objective Bayesian Edge Screening and Structure Selection for Networks of Binary Variables

Marsman, M.[1], Huth, K.[1,2], Waldorp, L. J.[1] & Ntzoufras, I.[3]
1 University of Amsterdam
2 Centre for Urban Mental Health
3 Athens University of Economics and Business

Correspondence concerning this article should be addressed to:
Maarten Marsman
University of Amsterdam, Psychological Methods
Nieuwe Achtergracht 129B
PO Box 15906, 1001 NK Amsterdam, The Netherlands
E-mail may be sent to m.marsman@uva.nl.

MM was supported by a Veni grant (451-17-017) from the Netherlands Organization for Scientific Research (NWO).

Abstract

The Ising model is one of the most widely analyzed graphical models in network psychometrics. Unfortunately, popular approaches to parameter estimation and structure selection for the Ising model cannot naturally express uncertainty about the estimated parameters or selected structures. To address this issue, this paper offers an objective Bayesian approach to parameter estimation and structure selection for the Ising model. Our approach builds on George and McCulloch's continuous spike-and-slab approach (1993, Journal of the American Statistical Association, 88, 881–889). We show that our methods consistently select the correct structure and provide a new objective method to set the spike-and-slab hyperparameters. To circumvent the exploration of the complete structure space, which is too large in practical situations, we propose a novel approach that first screens for promising edges and then only explore the space instantiated by these edges. We apply our proposed methods to estimate the network of depression and alcohol use disorder symptoms from symptom scores of over 26,000 subjects.

Keywords: Bayesian, Ising model, spike and slab, depression, alcohol use disorder

Introduction

Undirected graphical models, also known as Markov random fields (MRFs; Kindermann & Snell, 1980), have become an indispensable tool to describe the complex interplay of variables in many fields of science. The Ising model (Ising, 1925), or quadratic exponential model (Cox, 1972), is one MRF that attracted the interest of psychologists. It is defined by the following probability distribution over the configurations of a p-dimensional vector $\mathbf{x}$, with $\mathbf{x} \in \{0, 1\}^p$,

$$p(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\mathrm{Z}(\boldsymbol{\mu}, \boldsymbol{\Sigma})} \exp\left(\sum_{i=1}^{p} x_i \mu_i + \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} x_i x_j \sigma_{ij} \right), \tag{1}$$

which covers all main effects μ_i and pairwise associations σ_{ij} of the p binary variables. The pairwise associations encode the conditional dependence and independence relations between variables in the model: If an association is equal to zero, the two variables are independent given the rest of the variables, and there is no direct relation between them. Otherwise, the two variables are directly related. These relations can be visualized as edges in a network, where the model's variables populate the network's nodes. This view of the Ising model in psychological applications inspired the field of network psychometrics (Epskamp, Maris, Waldorp, & Borsboom, 2018; Marsman et al., 2018), which now spans research in, among others, personality (Constantini et al., 2019; Cramer et al., 2012), psychopathology (Borsboom & Cramer, 2013; Cramer et al., 2016), attitudes (Dalege et al., 2016; Dalege, Borsboom, van Harreveld, & van der Maas, 2019), educational measurement (Marsman, Maris, Bechger, & Glas, 2015; Marsman, Tanis, Bechger, & Waldorp, 2019), and intelligence (Savi, Marsman, van der Maas, & Maris, 2019; van der Maas, Kan, Marsman, & Stevenson, 2017).

The primary objective in empirical applications of the Ising model is determining the network's structure or topology. Three practical challenges complicate this objective. The first practical challenge is the normalizing constant $\mathrm{Z}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ in Eq. (1), which is a sum over all 2^p possible configurations of the binary vector $\mathbf{x}$. Even for small graphs, this normalizing constant can be expensive to compute. For example, for a network of 20 variables, the normalizing constant consists of more than one million terms. Given that the normalizing constant is repeatedly evaluated in numerical optimization or simulation approaches to estimate the model's parameters, the direct computation of the likelihood is computationally intractable. The second practical challenge in determining the Ising model's structure is the balance between model complexity and data. With p main effects and $\binom{p}{2}$ pairwise interactions, the number of free parameters can quickly overwhelm the limited information in available data. The third practical challenge is the efficient selection of a structure with desirable statistical properties from the vast space of possible structures. For a network of 20 variables, the structure space comprises $2^{190} \approx 1e^{57}$ potential structures, which is simply too large to enumerate in practice.

In psychology, eLasso is the structure selection solution for the Ising model and overcomes all three challenges. First, it adopts a pseudolikelihood approach to circumvent the normalizing constant. The pseudolikelihood replaces the joint distribution of the vector variable $\mathbf{x}$ —i.e., the full Ising model in Eq. (1)— with its respective full-conditional distributions:

$$p^*(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \prod_{i=1}^{p} p(x_i \mid \mathbf{x}^{(i)}, \mu_i, \boldsymbol{\sigma}_i^{(i)}) = \frac{\exp\left(\sum_{i=1}^{p} x_i \mu_i + \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} x_i x_j \sigma_{ij} \right)}{\prod_{i=1}^{p} \left(1 + \exp\left(\mu_i + \sum_{j \neq i} \sigma_{ij} x_j \right) \right)}, \tag{2}$$

where $\boldsymbol{\sigma}_i^{(i)} = (\sigma_{i1}, \ldots, \sigma_{i(i-1)}, \sigma_{i(i+1)}, \ldots \sigma_{ip})^\mathsf{T}$. Observe that the pseudolikelihood is equivalent to Eq. (1) except that it replaces the intractable normalizing constant with a tractable one. Second, eLasso balances structure complexity with the information available from the data at hand using the Lasso (Tibshirani, 1996). With Lasso estimation, eLasso imposes the l_1-penalty on the pseudolikelihood parameters,

$$\min_{\mu_i,\, \boldsymbol{\sigma}_i^{(i)}} \left\{ \ln p^*(x_i \mid \mu_i, \boldsymbol{\sigma}_i^{(i)}) \right\}, \text{ subject to } \sum_{j \neq i} |\sigma_{ij}| \leq \rho,$$

effectively shrinking negligible effects to precisely zero. In eLasso, the l_1-penalty is imposed on the parameters of each full-conditional separately, and the full-conditionals are optimized one at a time. Meinshausen and Bühlmann (2006) showed that the pseudolikelihood in combination with Lasso can consistently uncover the true topology for normally distributed data, and Ravikumar, Wainwright, and Lafferty (2010) showed this in case of the Ising model. Third, eLasso selects the structure that optimizes the parameters subject to the l_1 constraint, which is specified up to its tuning parameter ρ. It performs the optimization for a range of values for the tuning parameter and then selects the value that minimizes an extended Bayesian information criterion (Barber & Drton, 2015; Chen & Chen, 2008). Thus, structure selection with eLasso is analogous to selecting the tuning parameter. Since a distinct tuning parameter is imposed on each full-conditional, which are optimized in isolation, results might differ across the full-conditionals. For example, an association is estimated to be precisely zero in one full-conditional but nonzero in another. Currently, eLasso uses two ad-hoc rules to combine these results into a single structure: The AND-rule includes an edge between two variables if both estimates of their association are nonzero, and the OR-rule includes an edge between two variables if either estimate is nonzero. Even though they are ad-hoc, both rules can consistently uncover the true underlying topology. This combination of methods allows eLasso to efficiently perform structure selection for the Ising model, which is why it has become widely popular in psychometric practice.

We, however, have three concerns with eLasso. Our first concern is that eLasso cannot express the uncertainty associated with a selected structure, and thus does not inform us about other plausible structures for the data at hand. A structure's plausibility is disclosed in its posterior probability. To compute the posterior probability we have to entertain multiple structures and take their prior plausibility into account, but eLasso searches for a single optimal structure instead. Our second concern is that the aforementioned ad-hoc rules that are used to decide if edges are in the network do not weigh the available evidence in a balanced way. The full-conditionals in Eq. (2) are asymmetric, even though the associations are not, suggesting that the two ad-hoc rules are suboptimal from a statistical viewpoint. Critics of eLasso often point to this asymmetry (e.g., Brusco, Steinley, Hoffman, Davis-Stober, & Wasserman, in press). Our third concern is that eLasso does not articulate the precision of the parameters it estimates. There are two underlying causes for this. First, eLasso produces two estimates for every association parameter, one for every full-conditional it is in. ELasso combines these estimates into a single number, but it is unclear how this impacts parameter uncertainty. This problem is exacerbated by the asymmetry of the full-conditionals, and thus the estimates that come from them. Second, standard expressions for parameter uncertainty are unavailable for Lasso estimation (Tibshirani, 1996). As a result, it is unclear what the best way is to communicate uncertainty in Lasso parameter estimates (e.g., Kyung, Gill, Ghosh, & Casella, 2010). Since the limiting distribution of the Lasso estimator is non-Gaussian with a point mass at zero (e.g., Knight & Fu, 2000; Pötscher & Leeb, 2009), standard solutions such as the bootstrap, although frequently used (see, for instance, Epskamp, Borsboom, & Fried, 2018; Tibshirani, 1996), cannot be

used to obtain confidence intervals or standard errors (e.g., Bühlmann, Kalisch, and Meier, 2014, Section 3.1; Pötscher and Leeb, 2009). It appears that Bayesian formulations of the Lasso offer a more natural framework for uncertainty quantification (Kyung et al., 2010; Park & Casella, 2008), but approximate confidence intervals/standard errors could also be obtained by desparsifying the Lasso (Bühlmann et al., 2014; van de Geer, Bülmann, Ritov, & Dezeure, 2014)

In light of these concerns, our goals are threefold. Our primary goal is to introduce a Bayesian alternative to eLasso for learning the topology of Ising models. Bayesian approaches to model selection often introduce binary indicators γ for the selection of variables in the model (e.g., George & McCulloch, 1993; O'Hara & Sillanpää, 2009). We will use these indicators here to model edge selection: If the indicator γ_{ij} equals one, the edge between variables i and j is included. Otherwise, the edge is excluded. A structure s is then a specific configuration of a vector of $\binom{p}{2}$ indicator variables $\boldsymbol{\gamma}_s$, and the collection of network structures is equal to

$$\mathcal{S} = \{0, 1\}^{\binom{p}{2}}.$$

We wish to estimate the posterior structure probabilities $p(\boldsymbol{\gamma} \mid \mathbf{x})$, since they convey all the information that is available on the structures $\boldsymbol{\gamma} \in \mathcal{S}$, and can be used to express the plausibility of a particular structure or the inclusion of a specific edge for the data at hand. To unlock these Bayesian benefits (see Marsman & Wagenmakers, 2017; Wagenmakers, Marsman, et al., 2018, for detailed examples), we have to connect the indicator variables to the selection problem at hand.

Our secondary goal is to formulate a continuous spike-and-slab approach, initially proposed by George and McCulloch (1993) for covariate selection in regression models, for edge selection in networks. In this approach, the binary indicators are used to hierarchically model the prior distributions of focal parameters by assigning zero-centered diffuse priors to effects that should be included and priors that are sharply peaked about zero to negligible effects. These continuous spike-and-slab components are usually Gaussian (e.g., George & McCulloch, 1993; Ročková & George, 2014) or Laplace distributions (e.g., Ročková, 2018; Ročková & George, 2018). Even though the Laplace distribution generates a Bayesian Lasso (Park & Casella, 2008), its drawback is that its posterior distribution is difficult to approximate using computational tools other than simulation. We therefore adopt Gaussian spike-and-slab components in our edge selection approach.

Our tertiary goal is to analyze the full or joint pseudolikelihood in Eq. (2) instead of analyzing the full-conditionals in isolation. Analyzing the joint pseudolikelihood offers two advantages. The first advantage is that it allows us to circumvent the use of the ad-hoc rules in structure selection. The second advantage is that it allows us to formulate a single prior distribution for the focal parameters, such that we obtain a single posterior distribution, which we can analyze in a meaningful way. In sum, by adopting the joint pseudolikelihood we can weigh the available information for structure selection and parameter estimation in a balanced way. The disadvantage of using the joint pseudolikelihood is its increased computational expense for some of the numerical procedures, and the inability of analyzing the full-conditionals in parallel. However, for the network sizes that are typically encountered in psychological applications this increase in computational expense is negligible.

The continuous spike-and-slab approach to select a network's topology poses three critical challenges that we address in this paper. The first challenge that we address is the consistency of the structure selection procedure. In a recent analysis of covariate selection in linear regression, Narisetty and He (2014) showed that the continuous spike-and-slab approach is inconsistent

if the hyperparameters are not correctly scaled. We extend this observation to the current structure selection problem and prove that a correct scaling of the hyperparameters leads to a consistent structure selection approach in an embedding with p fixed, n increasing. The second challenge that we address is the specification of tuning parameters. The effectiveness of the continuous spike-and-slab set-up crucially depends on their specification. Unfortunately, objective methods to specify these parameters are currently unavailable, and tuning them is difficult and context dependent (e.g., George & McCulloch, 1997; O'Hara & Sillanpää, 2009). To overcome this issue, we develop a new procedure to automatically set the tuning parameters in such a way that we achieve a high specificity, similar to the performance of eLasso. Other approaches are considered in the Discussion. The final challenge that we address is the practical exploration of the structure space $\mathcal{S}$. Even for relatively small networks, the structure space $\mathcal{S}$ can be vast, and exploring it poses a significant challenge. Moreover, the posterior distribution over the structure space may be diluted (George, 1999), which implies that even the most plausible structures have relatively small posterior probabilities and many similar structures exist. As a result, valuable computational effort is wasted on relatively uninteresting structures and it is difficult to estimate their probabilities with reasonable precision. To overcome this issue, we propose a novel, two-step approach. We first employ a deterministic estimation approach (Ročková & George, 2014), utilizing an expectation-maximization (EM; Dempster, Laird, & Rubin, 1977) variant of the continuous spike-and-slab approach to screen for a subset of promising edges. We then use a stochastic estimation approach (George & McCulloch, 1993), utilizing a Gibbs sampling (Geman & Geman, 1984) variant to explore the structure space instantiated by these promising edges. In sum, we propose a coherent Bayesian methodology for structure selection for the Ising model. The freely available R package `rbinnet` implements the proposed methods.[1]

The remainder of this paper is organized as follows. After this introduction, we first specify our Bayesian model, i.e., we discuss the pseudolikelihood and prior set-up. Then, we analyze the consistency of our spike-and-slab approach for structure selection and show that it is consistent if suitably scaled. We wrap up the blueprint of our Bayesian model with the objective specification of hyperparameters for our spike-and-slab set-up. We then present an EM and a Gibbs implementation of our Bayesian structure selection set-up used for edge screening and structure selection, respectively. In our suite of Bayesian tools, edge screening most closely resembles eLasso, and we will compare the performance of these two methods in a series of simulations. Finally, we present a full analysis of data on alcohol abuse and major depressive disorders from the National Survey on Drug Use and Health. As far as we know, these two disorders have not been analyzed on a symptom level together in a network approach.

Bayesian Model Specification

The set-up of any Bayesian model comprises two parts: The likelihood of the model's parameters and their prior distributions. We start with the likelihood dictated by the Ising model, and the pseudolikelihood approach that we adopt to circumvent the computational intractability of the full Ising model. We follow up with the specification of prior distributions for the Ising model's parameters, tying in George and McCulloch's 1993 continuous spike-and-slab prior set-up for edge selection.

[1] The package can be downloaded from `https://github.com/MaartenMarsman/rbinnet/`

The Ising Model Pseudolikelihood

In this paper, we will adopt the pseudolikelihood approach of Besag (1975), as presented in Eq. (2). We will furthermore assume that the observations are independent and identically distributed, such that the full pseudolikelihood becomes

$$p^*(\mathbf{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{v=1}^{n} p^*(\mathbf{x}_v \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where $\mathbf{X} = (\mathbf{x}_1^\intercal, \ldots, \mathbf{x}_n^\intercal)^\intercal$, and we have adopted v to index the n independent and identically distributed observations. Both maximum pseudolikelihood and Bayesian pseudoposterior estimates are consistent as n increases (e.g., Arnold & Strauss, 1991; Geys, Molenberghs, & Ryan, 2007; Miller, 2019), and can consistently uncover the unknown graph structure of the full Ising model (Barber & Drton, 2015; Csiszár & Talata, 2006; Meinshausen & Bühlmann, 2006; Pensar, Nyman, Niiranen, & Corander, 2017; Ravikumar et al., 2010). As a result, the pseudolikelihood has become an indispensable tool in the structure selection of Ising models.

The Continuous Spike-and-Slab Prior Set-Up

There are several ways to bring the indicator variables into our Bayesian model (e.g., Dellaportas, Forster, & Ntzoufras, 2002; George & McCulloch, 1993; Kuo & Mallick, 1998). O'Hara and Sillanpää (2009) and Consonni, Fouskakis, Liseo, and Ntzoufras (2018) provide two recent overviews. In this paper, we adopt the continuous spike and slab approach, which comprises two parts. First, a mixture of two zero-centered normal distributions is imposed on the focal parameters. Here, the focal parameters are the pairwise associations σ_{ij}. The indicator variables are then used to distinguish between the two mixture components, and thus the prior distribution on the focal parameters becomes

$$\sigma_{ij} \mid \gamma_{ij} \sim (1 - \gamma_{ij})\, \mathcal{N}(0, \nu_0) + \gamma_{ij}\, \mathcal{N}(0, \nu_1),$$

where $\mathcal{N}(0, \nu)$ denotes the normal distribution with a mean equal to zero and a variance equal to ν. A small but positive variance $\nu_0 > 0$ is assigned to the component that is associated to $\gamma_{ij} = 0$ to encourage the exclusion of negligible nonzero values, and a large variance $\nu_1 >> \nu_0$ is assigned to the component associated to $\gamma_{ij} = 1$ to accommodate all plausible values of the interaction. Second, the selection variables are *a priori* modeled as i.i.d. Bernoulli(θ) variables, which implies the following prior distribution on the structures $\boldsymbol{\gamma}_s$,

$$p(\boldsymbol{\gamma}_s) = \theta^{\gamma_{s++}} (1 - \theta)^{\binom{p}{2} - \gamma_{s++}}, \tag{3}$$

where $\gamma_{s++} = \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} \gamma_{sij}$, with $\gamma_{ij} = \gamma_{ji}$. Once the hyperparameters ν_0, ν_1 and θ are set, and the nuisance parameters are assigned a prior distribution, the posterior structure probabilities can then be estimated using, for example, a Gibbs sampler (Geman & Geman, 1984; George & McCulloch, 1993). We stipulate independent standard-normal prior distributions on the nuisance parameters $\boldsymbol{\mu}$, and make the objective specification of hyperparameters the topic of the ensuing sections.

Structure Selection Consistency

In this section, we analyze posterior selection consistency, the ability of our method to determine the correct network structure consistently. As alluded to in the introduction, selection consistency using George and McCulloch's spike and slab approach crucially depends on the hyperparameters ν_0 and ν_1. Unfortunately, fixing these parameters does not guarantee that our structure

selection procedure is consistent. Narisetty and He (2014) showed that the use of fixed constants may lead to an inconsistent selection procedure in the context of linear regression. Below, we will demonstrate that this is also the case in the context of structure selection for Ising models. However, we will also show that our selection approach is consistent if the spike variance ν_0 shrinks as a function of n. Narisetty and He presented a similar result for linear regression.

We first work out the concepts relevant for selection consistency, such as the posterior structure probability, and derive an approximate Bayes factor that is useful for the large-sample analysis. Then, we analyze the case with fixed hyperparameters and show that the selection procedure is inconsistent for fixed p, increasing n. Finally, we analyze the situation where the spike variance shrinks with n, and show that this shrinking hyperparameter set-up leads to a consistent selection procedure for fixed p, increasing n.

Selection consistency

We assume that the true structure t is in the set $\mathcal{S}$.[2] We quantify our uncertainty in selecting a structure s, $s \in \mathcal{S}$, using the posterior structure probability

$$p(\boldsymbol{\gamma}_s \mid \boldsymbol{X}) = \frac{p^*(\mathbf{X} \mid \boldsymbol{\gamma}_s)\, p(\boldsymbol{\gamma}_s)}{\sum_{s \in \mathcal{S}} p^*(\mathbf{X} \mid \boldsymbol{\gamma}_s)\, p(\boldsymbol{\gamma}_s)} = \frac{\mathrm{BF}^*_{st}\, \mathrm{o}_{st}}{1 + \sum_{u \in \mathcal{S}_{\setminus t}} \mathrm{BF}^*_{ut}\, \mathrm{o}_{ut}},$$

where $p^*(\mathbf{X} \mid \boldsymbol{\gamma}_s)$ denotes the integrated pseudolikelihood for the structure s, BF^*_{st} the Bayes factor pitting structure s against the correct structure t, and o_{st} denotes the prior model odds of the two structures. Selection consistency requires us to show that the posterior structure probabilities $p(\boldsymbol{\gamma}_s \mid \mathbf{X})$ tend to zero for structures $s \neq t$, and that $p(\boldsymbol{\gamma}_t \mid \mathbf{X})$ tends to one as the sample size grows. This is equivalent to showing that the Bayes factors BF_{st} tend to zero for structures $s \neq t$. Unfortunately, analytic expressions for the Bayes factors are currently unavailable. To come to a workable expression for the Bayes factor, we first redefine it in terms of the expected prior odds under the correct posterior distribution

$$\begin{aligned}
\mathrm{BF}^*_{st} &= \frac{p^*(\mathbf{X} \mid \boldsymbol{\gamma}_s)}{p^*(\mathbf{X} \mid \boldsymbol{\gamma}_t)} \\
&= \frac{\int\int p^*(\mathbf{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})\, p(\boldsymbol{\mu})\, p(\boldsymbol{\Sigma} \mid \boldsymbol{\gamma}_s)\, \mathrm{d}\boldsymbol{\Sigma}\, \mathrm{d}\boldsymbol{\mu}}{\int\int p^*(\mathbf{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})\, p(\boldsymbol{\mu})\, p(\boldsymbol{\Sigma} \mid \boldsymbol{\gamma}_t)\, \mathrm{d}\boldsymbol{\Sigma}\, \mathrm{d}\boldsymbol{\mu}} \\
&= \int\int \frac{p^*(\mathbf{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})\, p(\boldsymbol{\mu})\, p(\boldsymbol{\Sigma} \mid \boldsymbol{\gamma}_s)}{p^*(\mathbf{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})\, p(\boldsymbol{\mu})\, p(\boldsymbol{\Sigma} \mid \boldsymbol{\gamma}_t)} \\
&\quad \cdot \times \frac{p^*(\mathbf{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})\, p(\boldsymbol{\mu})\, p(\boldsymbol{\Sigma} \mid \boldsymbol{\gamma}_t)}{\int\int p^*(\mathbf{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})\, p(\boldsymbol{\mu})\, p(\boldsymbol{\Sigma} \mid \boldsymbol{\gamma}_t)\, \mathrm{d}\boldsymbol{\Sigma}\, \mathrm{d}\boldsymbol{\mu}}\, \mathrm{d}\boldsymbol{\Sigma}\, \mathrm{d}\boldsymbol{\mu} \\
&= \mathbb{E}^*\left(\frac{p(\boldsymbol{\Sigma} \mid \boldsymbol{\gamma}_s)}{p(\boldsymbol{\Sigma} \mid \boldsymbol{\gamma}_t)} \,\middle|\, \mathbf{X}, \boldsymbol{\gamma}_t \right) \\
&= \mathbb{E}^*\left(\prod_{i=1}^{p-1} \prod_{j=i+1}^{p} \frac{p(\sigma_{ij} \mid \gamma_{sij})}{p(\sigma_{ij} \mid \gamma_{tij})} \,\middle|\, \mathbf{X}, \boldsymbol{\gamma}_t \right),
\end{aligned}$$

[2] This $\mathcal{S}$-closed view might be unrealistic in practice. If the correct structure is not in $\mathcal{S}$, t is the structure closest in Kullback-Leibler divergence to the true structure.

which is the posterior expectation of the ratio of the prior distributions of $\boldsymbol{\Sigma}$ for the two models, s and t, under the correct structure specification $\boldsymbol{\gamma}_t$. This is a convenient representation, as we only have to consider the pseudoposterior distribution under the correct network structure. Observe that this representation also holds when the full Ising likelihood is used, except that in the latter case the Bayes factor BF_{st} is expressed as the expected prior odds w.r.t the posterior distribution and not the pseudoposterior distribution. For a fixed network of p variables, the posterior distribution can be accurately approximated with a normal distribution as n becomes large (see, for instance, Miller, 2019, Theorem 6.2), and the same holds for the pseudoposterior distribution (see, for instance, Miller, 2019, Theorems 3.2 and 7.3). To come to a workable expression of the Bayes factor we therefore approximate the pseudoposterior with a normal distribution, which leads to the following first order approximation of the Bayes factor (Tierney, Kass, & Kadane, 1989, Eq. 2.6),

$$\begin{aligned}
\text{BF}^*_{st} &= \prod_{i=1}^{p-1} \prod_{j=i+1}^{p} \frac{p(\hat{\sigma}_{ij} \mid \gamma_{sij})}{p(\hat{\sigma}_{ij} \mid \gamma_{tij})} \left[1 + O(n^{-1})\right] \\
&\approx \prod_{i=1}^{p-1} \prod_{j=i+1}^{p} \frac{p(\hat{\sigma}_{ij} \mid \gamma_{sij})}{p(\hat{\sigma}_{ij} \mid \gamma_{tij})} \\
&= \prod_{i=1}^{p-1} \prod_{j=i+1}^{p} \left(\sqrt{\frac{\nu_0}{\nu_1}} \exp\left(\hat{\sigma}^2_{ij} \frac{\nu_1 - \nu_0}{2\nu_1\nu_0} \right) \right)^{\gamma_{sij} - \gamma_{tij}}
\end{aligned} \tag{4}$$

where $\widehat{\boldsymbol{\Sigma}} = [\hat{\sigma}_{ij}]$ is the mode of $p^*(\boldsymbol{\Sigma}, \boldsymbol{\mu} \mid \mathbf{X}, \boldsymbol{\gamma}_t)$, or $p(\boldsymbol{\Sigma}, \boldsymbol{\mu} \mid \mathbf{X}, \boldsymbol{\gamma}_t)$ if the full Ising likelihood is used, and $O(n^{-1})$ denotes a rest term that is of order $1/n$. Since the pseudoposterior is consistent (c.f., Miller, 2019, Theorem 7.3), the Bayes factor using the pseudolikelihood and the full likelihood will converge to the same number.

We will show next that the approximate Bayes factors BF^*_{st}, for $s \neq t$, do not shrink to zero with the three hyperparameters fixed, but do shrink to zero if ν_0 shrinks to zero at a rate n^{-1}. The approximate Bayes factor comprises a product of the edge specific functions

$$f_{ij} = \left(\sqrt{\frac{\nu_0}{\nu_1}} \exp\left(\hat{\sigma}^2_{ij} \frac{\nu_1 - \nu_0}{2\nu_1\nu_0} \right) \right)^{\gamma_{s,ij} - \gamma_{t,ij}} \geq 0$$

which consists of two parts: The selection variables $\gamma_{s,ij}$ and $\gamma_{t,ij}$ that inform about the differences in edge composition of structures s and t, and the function $\sqrt{\nu_0/\nu_1} \exp\left(\hat{\sigma}_{ij}(\nu_1 - \nu_0)/2\nu_1\nu_0\right)$ that weighs in the contribution of the pseudoposterior. The edge specific function f_{ij} is equal to one if the edge is present in both structures, or is absent from both structures, since then $\gamma_{t,ij} - \gamma_{s,ij} = 0$. We therefore only have to consider what happens to the function f_{ij} for cases where $\gamma_{s,ij} \neq \gamma_{t,ij}$.

The Fixed Hyperparameter Case. If $\gamma_{t,ij}$ is equal to zero, and $\gamma_{s,ij}$ is equal to one, the correct value for the interaction parameter σ_{ij} is zero, and we observe that

$$f_{ij} \xrightarrow{n} \sqrt{\frac{\nu_0}{\nu_1}},$$

which, even though it is smaller than one and signals a preference for structure t, does not converge to zero as it should if the structure selection procedure would be consistent. If $\gamma_{t,ij}$ is equal to one, and $\gamma_{s,ij}$ is equal to zero, on the other hand, such that $|\sigma_{ij}| > 0$, we observe that

$$f_{ij} \xrightarrow{n} \sqrt{\frac{\nu_1}{\nu_0}} \exp\left(-\sigma^2_{ij} \frac{\nu_1 - \nu_0}{2\nu_1\nu_0} \right),$$

which does not converge to zero either. In fact, it may even signal a preference for the absence of the edge in structure s. These two observations indicate that the Bayes factors BF^*_{st} do not converge to zero, and thus the posterior probability $p(\boldsymbol{\gamma}_t \mid \mathbf{X})$ does not converge to one. In sum, the proposed structure selection procedure is inconsistent in the case that the three hyperparameters are fixed.

The Shrinking Hyperparameter Case. We next consider the case where ν_0 shrinks at a rate n^{-1}, and define $\nu_0 = \frac{\nu_1 \xi}{n}$. Here, ξ is a fixed (positive) penalty parameter that allows us some flexibility to emphasize the distinction between the spike and slab components. If, in this case, $\gamma_{t,\,ij}$ is equal to one and $\gamma_{s,\,ij}$ is equal to zero, the function f_{ij} is equal to

$$f_{ij} = \sqrt{\frac{n}{\xi}} \exp\left(-\hat{\sigma}^2_{ij} \frac{n-\xi}{2\nu_1 \xi}\right) = \exp\left(\frac{1}{2}\log\left(\frac{n}{\xi}\right) - \hat{\sigma}^2_{ij}\frac{n-\xi}{2\nu_1\xi}\right),$$

where the first factor tends to infinity, and the second factor tends to zero. Because the second factor tends to zero faster than the first factor tends to infinity, their product, again, tends to zero, as it should. On the other hand, if $\gamma_{t,\,ij}$ is equal to zero, the function f_{ij} becomes

$$\sqrt{\frac{\xi}{n}} \exp\left(\hat{\sigma}^2_{ij} \frac{n-\xi}{2\nu_1 \xi}\right),$$

where the first factor tends to zero, and the second factor tends to one because $\sqrt{n}\hat{\sigma}_{ij}$ tends to zero ($\hat{\sigma} = O_p(1/\sqrt{n})$). Therefore, f_{ij} tends to zero, as it should. In sum, the structure selection procedure is consistent if ν_0 shrinks at a rate of n^{-1}.

Objective Prior Specification

We follow the results in the previous section, and set the spike variance to $\nu_0 = \frac{\nu_1 \xi}{n}$, which leaves the specification of the slab variance ν_1, the penalty parameter ξ, and the prior inclusion probability to complete our Bayesian model blueprint. We first discuss a default setting for the spike and slab variances, i.e., the specification of ν_1 and ξ. We then discuss two options for the prior inclusion probabilities that we adopt in this paper.

Specification of the Spike and Slab Variances

One approach to find default values for the slab variance is to set it equal to n times the inverse of the Fisher information matrix $\mathcal{I}_\Sigma(\hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\mu}})^{-1}$, which approximately gives the information about σ_{ij} in a single observation, hence the name *unit information* (Kass & Wasserman, 1995). Kass and Wasserman (1995) showed that the logarithm of the Bayes factor —pitting one network structure against another— is approximately equal to the difference in Bayesian information criteria (BIC; Schwarz, 1978) of the two structures when we use unit information priors (see also, Raftery, 1999; Wagenmakers, 2007, for details). This result, combined with the fact that unit information priors can be automatically selected, makes them a popular approach in Bayesian variable selection. We follow the approach of Ntzoufras (2009), who achieved good results by setting the off-diagonal elements of $\mathcal{I}^{-1}$ to zero in the prior specification. This renders the spike-and-slab prior densities independent, and sets the slab variance to $\nu_{1,\,ij} = n\text{Var}(\hat{\sigma}_{ij})$.[3] If we set the slab variance equal to the

[3] We estimate $\text{Var}(\hat{\sigma}_{ij})$ by setting it to the diagonal of the Fisher information matrix for the pseudolikelihood, evaluated at the maximum pseudolikelihood estimate, c.f., Appendix A. Alternatively, one could approximate the variance by running a Gibbs sampling approach using non-informative priors. This, however, would take considerably more time in practice.

unit information the spike variance is equal to $\nu_{0,ij} = \xi \operatorname{Var}(\hat{\sigma}_{ij})$. Our structure selection procedure will still consistently select the correct structure, since $\nu_{0,ij}$ shrinks with rate n because $\operatorname{Var}(\hat{\sigma}_{ij})$ does (e.g., Miller, 2019, Section 5.2).

The spike-and-slab parameters are specified up to the constant ξ, which acts as a penalty parameter on the inclusion and exclusion of effects in the spike-and-slab prior. Larger values for ξ increases the overlap between the spike-and-slab components, and consequently makes it more likely that an effect is excluded, i.e., ends up in the spike component. It is the opposite case for smaller values. It is thus absolutely crucial to find a good value for this penalty. We wish to specify the tuning parameter ξ such that the performance of our edge selection approach is similar to that of eLasso. To that aim, we introduce an automated procedure to specify the tuning parameter such that the corresponding continuous spike-and-slab set-up is geared towards achieving a high specificity, or low type-1 error, similar to eLasso. The idea that we pursue here is to set the intersection of the spike-and-slab components equal to an approximate credible interval about zero. The left panel in Figure 1 illustrates the idea.

George and McCulloch (1993) show that the two densities intersect at

$$|\delta| = \sqrt{\nu_1 \frac{\log\left(\frac{\nu_1}{\nu_0}\right)}{\frac{\nu_1}{\nu_0} - 1}}.$$

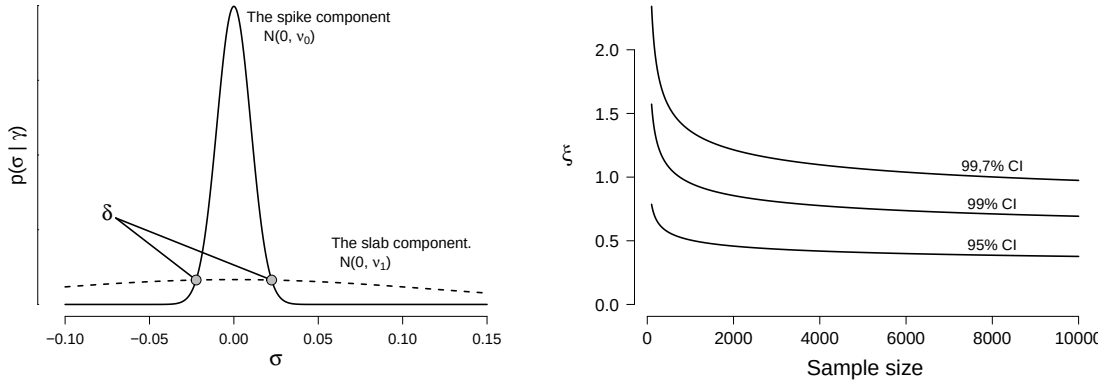
If we fill in our definitions for the spike and slab variances, the expression for $|\delta|$ boils down to.

$$|\delta| = \sqrt{n \operatorname{Var}(\hat{\sigma}_{ij}) \frac{\log\left(\frac{n}{\xi}\right)}{\frac{n}{\xi} - 1}}, \tag{5}$$

Where George and McCulloch (1993) discuss the subjective specification of δ, we explore its automatic specification by matching it to the approximate credible interval. We first determine the range of parameter values $(-|\delta|, |\delta|)$ considered to be insignificant, and then select the value of ξ such that the spike and slab components intersect at $\pm|\delta|$. When n is sufficiently large, the pseudoposterior distribution of an association parameter σ_{ij} is approximately normal (Miller, 2019), and $\operatorname{Var}(\hat{\sigma}_{ij})$ is its approximate variance. Thus, $(\hat{\sigma}_{ij} \pm 3\sqrt{\operatorname{Var}(\hat{\sigma}_{ij})})$ offers an approximate $99,7\%$ credible interval about the posterior mean $\hat{\sigma}_{ij}$. To set the variance of the spike distribution for negligible effects it is opportune to use the interval $(\pm 3\sqrt{\operatorname{Var}(\hat{\sigma}_{ij})})$, which offers an approximate credible interval about zero, i.e., the credible interval assuming that the edge i–j should, in fact, be excluded from the model. Equating the expression for $|\delta|$ on the right side of Eq. (5) with $3\sqrt{\operatorname{Var}(\hat{\sigma}_{ij})}$ gives:

$$\sqrt{n \frac{\log\left(\frac{n}{\xi}\right)}{\frac{n}{\xi} - 1}} = 3, \tag{6}$$

which we can solve numerically to obtain a value for ξ. Observe that, by specifying δ in this particular way, the penalty parameter ξ depends on the sample size but not the data or network's size. We denote the value of the penalty parameter that matches the intersection δ to the credible interval with ξ_δ. The relation between ξ_δ and sample size is illustrated in the right panel of Figure 1.



(a) Spike-and-slab prior.

(b) Relation between n and ξ_δ.

Figure 1. The left panel illustrates the spike-and-slab prior distribution and it's intersection point δ. The right panel illustrates the relationship between n and the ξ_δ value equating the intersection points $\pm\delta$ to three different credible intervals.

Specification of the Prior Inclusion Probability

Assuming that the correct structure is in $\mathcal{S}$, i.e., the $\mathcal{S}$-closed view of structure selection, a default choice to express ignorance or indifference between the structures in $\mathcal{S}$ is to stipulate a uniform prior distribution over the topologies in $\mathcal{S}$:

$$p(\boldsymbol{\gamma}_s) = \frac{1}{|\mathcal{S}|}, \text{ for } \boldsymbol{\gamma}_s \in \mathcal{S},$$

where $|\mathcal{S}|$ denotes the cardinality of the structure space. Here, the uniform prior is equal to

$$p(\boldsymbol{\gamma}_s) = 2^{-\frac{1}{2}p\,(p-1)},$$

and we can impose this prior on the structure space by fixing the prior inclusion probability θ in Eq. (3) to $\frac{1}{2}$. However, the uniform prior on the structure space does not take into account structural features of the models under consideration, such as sparsity. Various priors have been proposed as an alternative to accommodate these features (see Consonni et al., 2018, Section 3.6, for a detailed discussion). One particular issue inherent in structure comparisons is multiplicity, and Scott and Berger (2010) argue that the prior distribution should account for this. Consonni et al. (2018) show that stipulating a hyperprior on the prior inclusion probability θ accounts for multiplicity. In particular, they showed that the uniform hyperprior Beta(1, 1) leads to the following prior on the structure space

$$p(\boldsymbol{\gamma}_s) = p(\boldsymbol{\gamma}_s \mid \gamma_{s,++} = c) \times p(\gamma_{s,++} = c) = \frac{1}{\binom{p}{2} + 1} \times \frac{1}{\binom{\binom{p}{2}}{\gamma_{s,++}}},$$

where c denotes the complexity of structures, with $c \in (0, 1, \ldots, \binom{p}{2})$, i.e., the number of edges in the topology. Thus, instead of a uniform prior on the structure space, the hierarchical prior stipulates a uniform prior on the structure's complexity. As a result, it favors models that have a relatively extreme level of complexity, e.g., are densely connected or are sparsely connected.

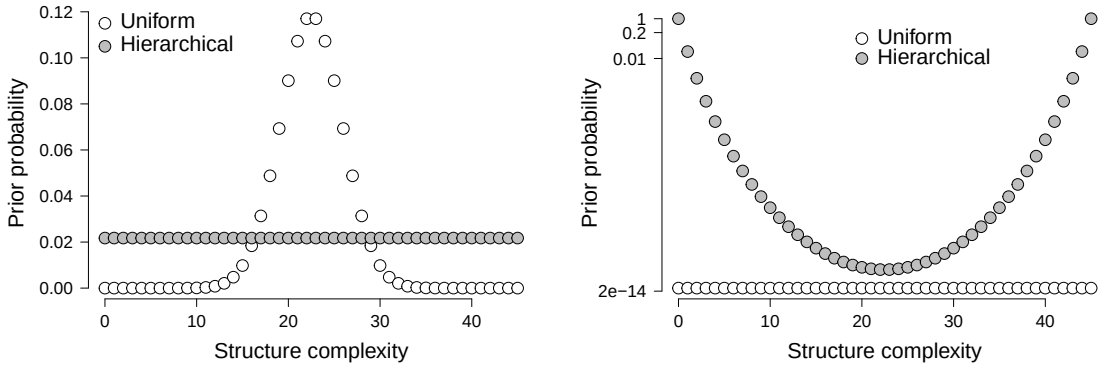


Figure 2. The left panel illustrates how the two prior distributions assign probabilities to structure complexity. The right panel illustrates how the the two prior distributions assign probabilities to different structures with the same complexity. The prior probabilities in the right panel are shown on a log scale. For both panels, $p = 10$.

Figure 2 illustrates the different probabilities that the two distributions assign to structure complexity, a priori, and the probabilities they assign to structures that have the same complexity (shown on a log scale). The left panel of Figure 2 shows that whereas the hierarchical prior is uniform on the complexity, the uniform prior is not, and favors structures that have approximately half of the available edges. However, the right panel of Figure 2 illustrates that the hierarchical prior emphasizes structures at the extremes of complexity. We will adopt both the uniform prior distribution on the structure space and the uniform prior distribution on the structure's complexity, and analyze them further in the section on numerical illustrations. Based on Figure 2, however, we expect that for small samples, the hierarchical prior will place much emphasis on extremely sparse structures, since our penalty selection approach already gears toward sparse solutions.

Bayesian Edge Screening and Structure Selection for the Ising model

George and McCulloch (1993) proposed stochastic search variable selection (SSVS) as a principled approach to Bayesian variable selection. SSVS uses the spike-and-slab prior specification to emphasize the posterior probability of promising structures and Gibbs sampling to extract this information from the data at hand. The Gibbs sampler is a powerful tool for the exploration of the posterior distribution of potential network structures. However, since the structure space $\mathcal{S}$ can be quite large in practical settings, it might take a while to sufficiently explore the posterior distribution and produce reliable estimates of the posterior structure probabilities. We, therefore, wish to prune the structure space by selecting the promising edges before running the Gibbs sampler. We explore an EM variable selection approach for this initial edge screening, and then follow-up with an SSVS approach for structure selection on the set of promising edges.

Edge Screening with EM Variable Selection

Ročková and George (2014) were the first to propose the use of EM for Bayesian variable selection, in combination with the spike-and-slab prior specification of George and McCulloch (1993), to covariate selection of linear models. The EM algorithm aims to find the posterior mode of the

pseudoposterior distribution $p^*(\boldsymbol{\Sigma}, \boldsymbol{\mu}, \theta \mid \mathbf{X})$ and does this by iteratively maximizing the "complete data" pseudoposterior distribution $p^*(\boldsymbol{\Sigma}, \boldsymbol{\mu}, \theta, \boldsymbol{\gamma} \mid \mathbf{X})$, treating the selection variables $\boldsymbol{\gamma}$ as missing or latent variables. The algorithm alternates between two steps. In the expectation or E-step we compute the expected log-pseudoposterior distribution, or Q-function,

$$\mathrm{Q}\left(\boldsymbol{\Sigma}, \boldsymbol{\mu}, \theta \mid \boldsymbol{\Sigma}^k, \theta^k\right) = \mathbb{E}\left(\ln p^*(\boldsymbol{\Sigma}, \boldsymbol{\mu}, \theta, \boldsymbol{\gamma} \mid \mathbf{X}) \,\middle|\, \boldsymbol{\Sigma}^k, \theta^k\right),$$

with respect to posterior distribution of the latent variables $p(\boldsymbol{\gamma} \mid \boldsymbol{\Sigma}^k, \theta^k)$, where $\boldsymbol{\Sigma}^k$, and θ^k denote the estimates in iteration k. The E-step is followed by a maximization or M-step in which we find the values $\boldsymbol{\Sigma}^{k+1}$, $\boldsymbol{\mu}^{k+1}$ and θ^{k+1} that maximize the Q-function. The two steps are repeated until convergence.

The E-step of the EM algorithm involves expectations of the latent or missing variables, i.e., the vector of selection variables $\boldsymbol{\gamma}$. Since the latent selection variables only operate in the spike-and-slab prior distributions, the derivation of the E-step will follow the derivation of Ročková and George (2014). For a complete treatment of EMVS, however, we include an analysis of both the E-step and the M-step in Appendix A. Appendix A also includes details about estimating the (asymptotic) posterior standard deviations from the EM output.

Edge Screening. The EM algorithm that we outlined in the previous section identifies a posterior mode $(\widehat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\mu}}, \hat{\theta})$, and we threshold the modal estimates to obtain a tightly matching network structure $\hat{\boldsymbol{\gamma}}$. The idea of Ročková and George (2014) that we pursue here is that "large" interaction effect estimates define a set of promising edges, and we can thus prune edges that link to "small" interaction effect estimates. We define the structure $\hat{\boldsymbol{\gamma}}$ that closely matches the modal estimates $(\widehat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\mu}}, \hat{\theta})$ to be the most probable structure $\boldsymbol{\gamma}$ given the parameter values $(\boldsymbol{\Sigma}, \boldsymbol{\mu}, \theta) = (\widehat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\mu}}, \hat{\theta})$, i.e.,

$$\hat{\boldsymbol{\gamma}} = \arg\max_{\boldsymbol{\gamma}} \; p(\boldsymbol{\gamma} \mid \widehat{\boldsymbol{\Sigma}}, \hat{\theta}). \tag{7}$$

For our Bayesian model, the posterior inclusion probabilities for the different edges are conditionally independent, and the posterior inclusion probability for an edge $i - j$ is given by

$$p(\gamma_{ij} \mid \hat{\sigma}_{ij}, \hat{\theta}) = \frac{p(\hat{\sigma}_{ij} \mid \gamma_{ij})\, p(\gamma_{ij} \mid \hat{\theta})}{\sum_{\Gamma_{ij}=\gamma_{ij}} p(\hat{\sigma}_{ij} \mid \gamma_{ij})\, p(\gamma_{ij} \mid \hat{\theta})}.$$

Thus, we obtain $\hat{\boldsymbol{\gamma}}$ from maximizing the inclusion and exclusion probabilities in Eq. (7), for each of the $\binom{p}{2}$ edges, which means that

$$\hat{\gamma}_{ij} = 1 \iff p(\gamma_{ij} = 1 \mid \hat{\sigma}_{ij}, \hat{\theta}) \geq 0.5,$$

and we prune the edges for which $p(\gamma_{ij} = 0 \mid \hat{\sigma}_{ij}, \hat{\theta}) \geq 0.5$. This edge selection and pruning approach leads to a structure $\boldsymbol{\gamma}$ that is a median probability model, as defined by Barbieri and Berger (2004) to be the structure comprising edges that have a posterior inclusion probability at or above a half.[4] Ročková and George (2014) show that instead of selecting the structure $\hat{\boldsymbol{\gamma}}$ based on

[4]With the caveat that Barbieri and Berger (2004) define the posterior inclusion probability of an edge *i–j* to be the aggregate of the posterior probabilities of structures that include the edge

$$p(\gamma_{ij} = 1 \mid \mathbf{X}) = \sum_{\boldsymbol{\gamma}_s : \gamma_{ij}=1} p(\boldsymbol{\gamma}_s \mid \mathbf{X})$$

where we define the posterior inclusion probabilities locally.

the posterior inclusion probabilities, we may equivalently select it through thresholding the values of $\hat{\sigma}_{ij}$. Specifically,

$$\hat{\gamma}_{ij} = 1 \iff |\hat{\sigma}_{ij}| \geq \sqrt{2 \log\left(\frac{1-\hat{\theta}}{\hat{\theta}} \sqrt{\frac{\nu_1}{\nu_0}}\right) \frac{\nu_1 \nu_0}{\nu_1 - \nu_0}} = \sqrt{2 \log\left(\frac{1-\hat{\theta}}{\hat{\theta}} \sqrt{\frac{n}{\xi}}\right) \frac{n \mathrm{Var}(\hat{\sigma}_{ij}) \xi}{n - \xi}}. \quad (8)$$

Such a connection between the magnitude of the modal estimates $\hat{\sigma}_{ij}$ and promising edges $i - j$ we envisioned from the beginning. Observe that, since $n\mathrm{Var}(\hat{\sigma}_{ij})$ is the unit information, i.e., it is a constant, the right-most factor shrinks with n. Moreover, it shrinks much faster than that $\log(\sqrt{n})$ tends to infinity, such that the threshold moves to smaller values as n increase, as it should.

Structure Selection with SSVS

The EMVS approach enables us to screen for a promising set of edges by locating a local posterior mode and pruning edges associated with small modal parameters. The structure $\boldsymbol{\gamma}'$ that comes out of this pruned edge set is a local median probability structure. We now wish to directly explore $p^*(\boldsymbol{\gamma} \mid \mathbf{X})$, the pseudoposterior distribution of network structures, to find out if $\boldsymbol{\gamma}'$ is also the global median probability model, and if there are other promising structures for the data at hand. We do this using the stochastic search and variable selection (SSVS) approach of George and McCulloch (1993), which essentially combines the spike-and-slab prior set-up with Gibbs sampling to produce a sequence

$$\boldsymbol{\gamma}^{(0)}, \boldsymbol{\gamma}^{(1)}, \boldsymbol{\gamma}^{(2)}, \ldots,$$

which converges in distribution to samples from $\boldsymbol{\gamma} \sim p(\boldsymbol{\gamma} \mid \mathbf{X})$. We then shift our focus to structures $\boldsymbol{\gamma}_s$ that occur frequently in the generated sequence, which are the structures that have a high posterior probability. We cut down the potentially large number of network structures that the Gibbs sampler needs to explore by applying SSVS only to the edges screened by EMVS.

The Gibbs sampler operates by iteratively simulating values from the conditional distributions of (a subset of) the model parameters given the (other parameters and the) observed data. Unfortunately for us, the full-conditional distributions of our Bayesian model are not available in closed form, as the normal prior distributions that we have specified are not conjugate to the pseudolikelihood. However, since the pseudolikelihood comprises a sequence of logistic regressions, we can use the data-augmentation strategy that was proposed by Polson, Scott, and Windle (2013a) to facilitate a simple Gibbs sampling approach, with full-conditionals that are easy to sample from. A similar approach to the Ising model's pseudolikelihood was considered by Donner and Opper (2017). Here we extend this idea to SSVS for the Ising model.

Polson et al. (2013a) proposed an ingenious data augmentation strategy based on the identity

$$\frac{\left(e^{\psi}\right)^a}{\left(1 + e^{\psi}\right)^b} = \frac{1}{2^b} e^{(a-b/2)\psi} \int_{\mathbb{R}^+} e^{-\frac{1}{2}\omega\psi^2} \, p(\omega) \, \mathrm{d}\omega,$$

where $p(\omega)$ is a Pólya–Gamma distribution. A key aspect of this augmentation strategy is that it relates the logistic function of a parameter ψ on the left to something that is proportional to a normal distribution on the right. Since our prior distributions are all (conditionally) normal, and the normal distribution is its own conjugate, the data-augmented full-conditionals will all be normal.

To wit, applied to the pseudolikelihood in Eq. (2), we find

$$\prod_{i=1}^{p} p^*(x_i \mid \mathbf{x}^{(i)}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) =$$
$$\frac{1}{2^p} \prod_{i=1}^{p} \int_{\mathbb{R}_+} e^{[x_i - 1/2]\left[\mu_i + \sum_{j \neq i} \sigma_{ij} x_j\right] - \frac{1}{2}\omega_i \left[\mu_i + \sum_{j \neq i} \sigma_{ij} x_j\right]^2} p(\omega_i)\, \mathrm{d}\omega_i,$$

and with normal prior distributions for the pseudolikelihoods parameters we readily find normal full-conditional distributions when we condition on the augmented variables $\boldsymbol{\omega}$. Another important aspect of the augmentation strategy is that the conditional distribution of the augmented variables ω given the pseudolikelihood parameters and the observed data $\mathbf{X}$ is again a Pólya-Gamma distribution. Polson et al. (2013a) and Windle, Polson, and Scott (2014) provide efficient rejection algorithms to simulate from this distribution.

With the Pólya-Gamma augmentation strategy in place, the Gibbs sampler iterates between five steps, which are detailed in Appendix C. The Gibbs output allows us to estimate a number of important quantities. For example, the posterior structure probabilities can be estimated as

$$p(\boldsymbol{\gamma}_s \mid \mathbf{X}) \approx \frac{1}{R} \sum_{r=1}^{R} I(\boldsymbol{\gamma}^{(r)} = \boldsymbol{\gamma}_s),$$

where $I(\cdot)$ is an indicator function that is equal to one if its conditions are satisfied and equal to zero otherwise, and the (global) posterior inclusion probabilities as,

$$p(\gamma_{ij} = 1 \mid \mathbf{X}) \approx \frac{1}{R} \sum_{r=1}^{R} \gamma_{ij}^{(r)},$$

were $\boldsymbol{\gamma}^{(r)}$, for $r = 1, , \ldots, R$, denotes R iterates of the Gibbs sampler. In a similar way one can compute quantities related to the model-averaged posterior distribution of the model parameters, e.g.,

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \mathbf{X}) = \sum_{\boldsymbol{\gamma}_s \in \mathcal{S}} p(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{\gamma}_s, \boldsymbol{X}) p(\boldsymbol{\gamma}_s \mid \mathbf{X}),$$

or any of its marginals. In sum, the Gibbs sampler grants us the full Bayesian experience.

Numerical Illustrations

Edge Screening on Simulated Data with Sparse Topologies

The eLasso approach of van Borkulo et al. (2014) is the most popular method for analyzing Ising network models in psychology. We wish to find out how our EMVS approach stacks up against eLasso, and we, therefore, use the simulation set-up of van Borkulo et al. (2014) to compare both methods. Specifically, we focus on the simulations that lead to their Table 2, where an Erdős and Rényi (1960) model is used to generate underlying sparse topologies, and normal distributions are used to simulate the model parameters.[5] In these simulations, we vary π, the probability of gener-

[5]We follow the simulation set-up of van Borkulo et al. (2014) and set the association parameters σ_{ij} equal to $|Z_{ij}|$, where the Z_{ij} are independent normal random variables with mean zero and variance 0.25. The main effect parameters μ_i are set to $-|Z_i|$, where Z_i is an independent normal random variable with mean $\sigma_{i+} = \sum_{j=1}^{p} \sigma_{ij}/2$, with $\sigma_{ii} = 0$, and variance $\sigma_{i+}^2/36$.

ating an edge between two variables, p, the number of variables, and n, the number of observations and generate 100 datasets for each combination of values for π, p, and n.

We analyze the simulated datasets using eLasso, using the default settings implemented in the `IsingFit` program (van Borkulo, Epskamp, & Robitzsch, 2016), i.e., the AND-rule and an EBIC penalty equal to 0.25. We also analyze the simulated datasets using EMVS in combination with the ξ_δ method and the uniform and hierarchical specifications of the prior structure probabilities. We follow van Borkulo et al. (2014), and express the quality of the estimated solution using its sensitivity and specificity. Sensitivity is the proportion of present edges that are recovered by the method,

$$\text{SEN} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}},$$

i.e., the true positive rate. Specificity is equal to the proportion of absent edges that are correctly recovered,

$$\text{SPE} = \frac{\text{True negative}}{\text{True negative} + \text{False positive}},$$

i.e., the true negative rate. For eLasso, edge inclusion refers to a nonzero association estimate using the AND approach. For EMVS, it is taken to mean that the posterior inclusion probability exceeds 0.5.

Table 1 shows the result of these simulations for the eLasso method in the column labelled "eLasso", and are similar to the results reported in Table 2 in van Borkulo et al. (2014). The first thing to note about these results is that eLasso has a high true negative rate across all simulations. This was to be expected, as l_1-regularization gears towards edge exclusions, which is why it performs best in the sparse network settings considered here. Indeed, it's specificity goes down as the networks become more densely connected (i.e., larger values of π). The true positive rate of eLasso is significantly worse than its specificity, especially for the smaller sample sizes. However, the sensitivity increases with sample size, which underscores earlier results that a larger sample size helps overcome the prior shrinkage effect of the lasso (e.g., Epskamp, Kruis, & Marsman, 2017).

Next, we consider the performance of EMVS. The results for EMVS when the penalty ξ is set to ξ_δ, the penalty value for which the intersection of the spike and the slab components aligns with the $99,7\%$ approximate credible interval, c.f. Eq. (6), are shown in the columns labeled ξ_δ in Table 1. We analyzed the data using the uniform prior on the model space —ξ_δ (U)— and with the hierarchical model —ξ_δ (H). The first striking result is that the performance of the ξ_δ approach combined with a uniform prior on the structure space performs almost identical to eLasso, making it a valuable Bayesian alternative to the classical eLasso approach. Observe that the specificity equals the coverage probability of the credible interval for all but the smallest sample size. Thus, as one might expect, the coverage probability specified dictates the method's type-1 error or specificity. The hierarchical prior on the structure space leads to an improvement to the already high specificity. For the smaller sample sizes, however, the method's sensitivity is very low, suggesting that it is, perhaps, too conservative for settings with small sample sizes.

n	p		$\pi = 0.1$			$\pi = 0.2$			$\pi = 0.3$		
			eLasso	ξ_δ(U)	ξ_δ(H)	eLasso	ξ_δ(U)	ξ_δ(H)	eLasso	ξ_δ(U)	ξ_δ(H)
100	10	SEN	.264	.221	.044	.233	.251	.027	.218	.216	.032
		SPE	.997	.991	1.000	.994	.994	1.000	.993	.990	1.000
	20	SEN	.165	.240	.009	.171	.180	.004	.182	.114	.003
		SPE	.998	.991	1.000	.997	.992	1.000	.991	.993	1.000
	30	SEN	.151	.202	.002	.140	.118	.001	.142	.048	.000
		SPE	.999	.992	1.000	.995	.993	1.000	.979	.995	1.000
500	10	SEN	.557	.608	.484	.593	.575	.504	.595	.529	.462
		SPE	.997	.996	1.000	.992	.994	1.000	.989	.996	.999
	20	SEN	.519	.558	.455	.542	.497	.388	.550	.411	.268
		SPE	.998	.996	1.000	.990	.997	1.000	.972	.996	1.000
	30	SEN	.520	.526	.380	.489	.388	.265	.368	.196	.091
		SPE	.998	.996	1.000	.985	.996	1.000	.954	.998	1.000
1,000	10	SEN	.697	.730	.633	.675	.685	.639	.699	.681	.608
		SPE	.997	.997	1.000	.989	.996	1.000	.985	.996	.999
	20	SEN	.643	.680	.598	.676	.630	.565	.657	.545	.464
		SPE	.996	.996	1.000	.987	.997	1.000	.964	.997	.999
	30	SEN	.655	.645	.570	.635	.517	.449	.431	.298	.206
		SPE	.997	.997	1.000	.980	.997	1.000	.957	.997	1.000
2,000	10	SEN	.783	.811	.727	.759	.807	.738	.789	.770	.735
		SPE	.998	.997	1.000	.995	.995	0.999	.984	.996	1.000
	20	SEN	.740	.790	.715	.784	.738	.697	.765	.657	.623
		SPE	.997	.996	1.000	.985	.996	.999	.960	.997	.999
	30	SEN	.748	.761	.700	.738	.665	.609	.598	.441	.353
		SPE	.996	.996	1.000	.976	.997	1.000	.940	.997	1.000

Table 1
Sensitivity and specificity, as a measure of performance of eLasso and EMVS using either a uniform (U) or hierarchical prior (H), matching the spike and slab intersections to an approximate 99,7% credible interval.

Parameter Estimation on Simulated Data with Dense Topology

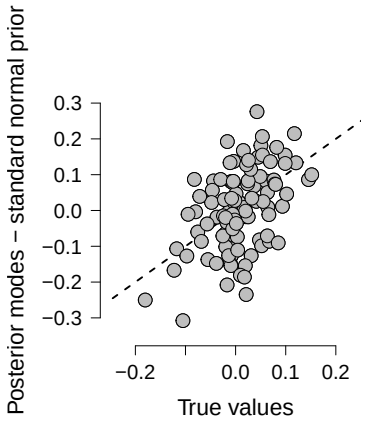
We continue with an illustration of the estimation of parameters and inclusion probabilities. For this analysis, we simulate data for $n = 20,000$ cases on a $p = 15$ variable network. The main effects were simulated from a Uniform$(-1, 1)$ distribution, and the matrix of associations $\mathbf{\Sigma}$ was set to $\mathbf{u}\mathbf{u}^\mathsf{T}$, where $\mathbf{u}$ is a p-dimensional vector of Uniform$(-\frac{1}{2}, \frac{1}{2})$ variables, such that the elements in $\mathbf{\Sigma}$ lie between $-\frac{1}{4}$ and $\frac{1}{4}$ and concentrate around zero. Observe that, in principle, this is a densely connected network as all edges have a nonzero value, although most effects will be very small and negligible. A second data set of $n = 2,000$ cases was used to compare the performance across different sample sizes.

Figure 3 shows the posterior mode estimates for the two sample sizes using a standard normal prior distribution in Panels (a) and (b) and using our spike-and-slab set-up, i.e., edge screening, in Panels (c) and (d). Observe that the effects are relatively small, and thus many observations are needed to retrieve reasonable estimates (Panels (a) and (b)). We, therefore, cull considerably more of the effects in the edge screening step for the smaller sample size than for the larger sample size (white dots indicate culled associations in Panels (c) and (d)). The horizontal gray lines in Panels (c) and (d) reveal the spike-and-slab intersections for the different associations (there are 210 different lines, which all lie very close to each other), the thresholds from Eq. (8). Effects that lie in between the two intersection points end up in the spike (not selected; white dots); otherwise, they end up in the slab (selected; gray dots). Note that the intersections points lie closer to each other for the larger sample size, as expected. Panels (e) and (f) show the maximum pseudolikelihood estimates for eLasso, subject to the l_1 constraint, which selects considerably fewer effects for the larger sample size, and a substantial shrinkage effect on the associations.[6]

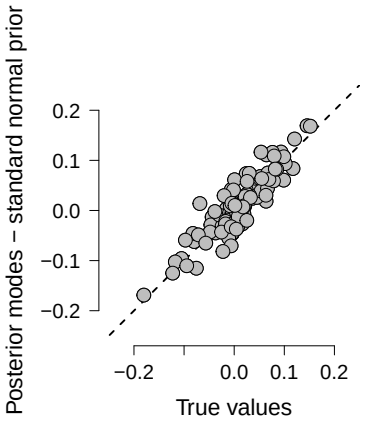
Figure 4 illustrates the various shrinkage effects in edge screening using EM and structure selection using the Gibbs sampler. Panels (a) and (b), for example, show that the procedures produce point estimates that are close to each other. Still, there is also variation between the two methods, especially around the spike and slab intersection lines. Although we did not show it here, the posterior estimates from EM and the Gibbs sampler were identical when we used a standard normal prior distribution instead of our spike-and-slab set-up. These observations suggest that the differences gleaned from Panels (a) and (b) come from the fact that the edge screening procedure optimizes the vector of inclusion variables with EM while the structure selection procedure averages over them in the Gibbs sampler. These differences become even more apparent when we compare the inclusion probabilities they estimate. Panels (c) and (d) show the inclusion probabilities against the posterior mode estimates for the edge screening approach, and Panels (e) and (f) show the inclusion probabilities against the posterior mean estimates for the structure selection procedure. Whereas the inclusion probabilities lie close to zero or one for the EM approach, they show a much smoother relation for the Gibbs sampling approach. The ability to estimate inclusion probabilities that are close to one or zero is called separation, and it is clear that the EM approach shows a better separation than the Gibbs approach. But the spike-and-slab Gibbs sampling approach, i.e., SSVS, already shows excellent separation compared to other methods (e.g., O'Hara & Sillanpää, 2009). Even though the edge screening approach shows better separation, it is also more liberal, as it includes more effects into the model than the structure selection procedure does. Panels (a) and (b) indicates these points in gray in Panels (a) and (b).

---

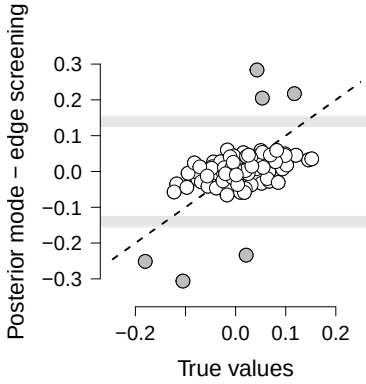
[6] In the sparse setting, one expects to detect approximately $\sqrt{\frac{n}{log(p(p-1)/2)}}$ edges with Lasso given a penalty of approximately $\sqrt{\frac{log(p(p-1)/2)}{n}}$. This translates to approximately 20 edges with 2,000 observations and 60 with 20,000 observations. In this simulation, we used a dense structure set-up, a set-up for which eLasso was not designed. The penalty values selected by eLasso could be used to analyze the performance of eLasso in practical situations, where the lower bound would be approximately $\sqrt{\frac{log(p(p-1)/2)}{n}}$. Here, the selected penalty values ranged between 0.0153 and 0.0319 with 2,000 observations. These values range below the lower bound of 0.0482. With 20,000 observations, we find values ranging between 0.0057 and 0.0158, which also largely range below the lower bound of 0.0153. This indicates that eLasso has difficulty with the non-sparse setting simulated here.



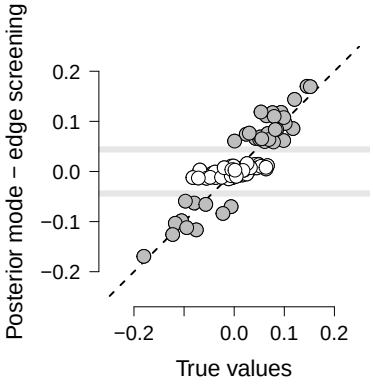
(a) Standard normal prior ($n = 2,000$).



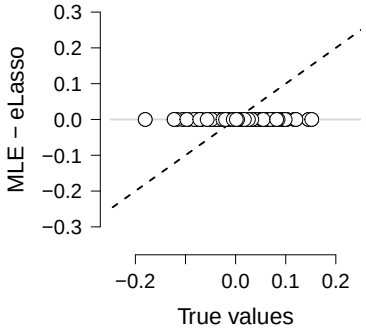
(b) Standard normal prior ($n = 20,000$).



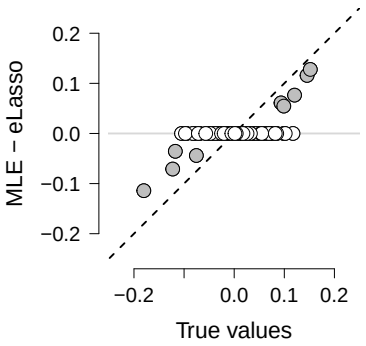
(c) Edge screening ($n = 2,000$).



(d) Edge screening ($n = 20,000$).

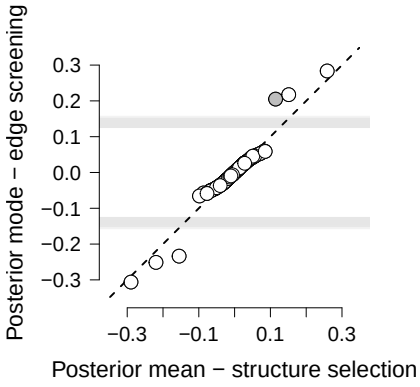


(e) ELasso ($n = 2,000$).

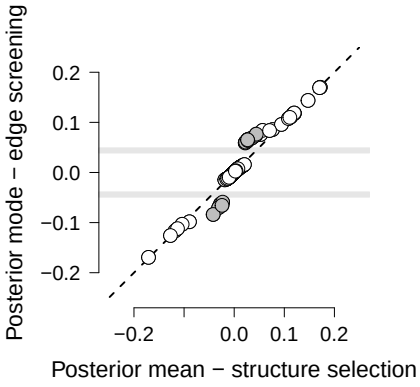


(f) ELasso ($n = 20,000$).

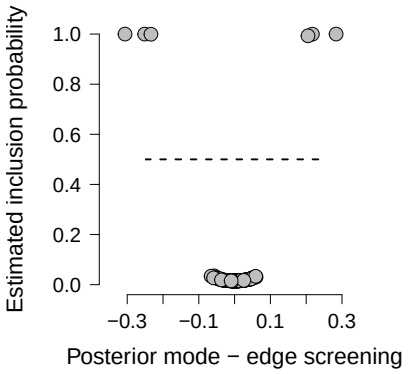
Figure 3. The posterior modes of the association parameters using a standard normal prior are shown in the top two panels, and the posterior modes of the associations using our spike-and-slab prior set-up, i.e., edge screening, are shown in the middle two panels. The horizontal gray lines in Panels (c) and (d) reveal the thresholds from Eq. (8). The bottom two panels show the maximum pseudolikelihood estimates produced by eLasso. The dashed lines are the bisection lines.



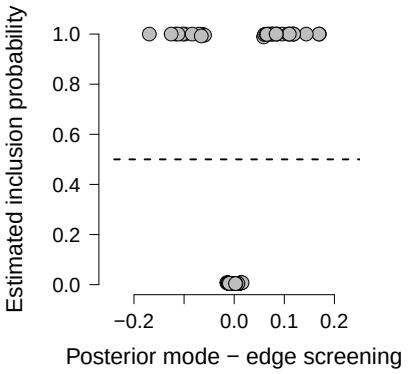
(a) Edge screening and structure selection estimates ($n = 2,000$).



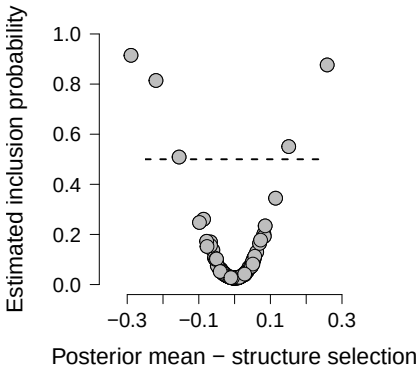
(b) Edge screening and structure selection estimates ($n = 20,000$).



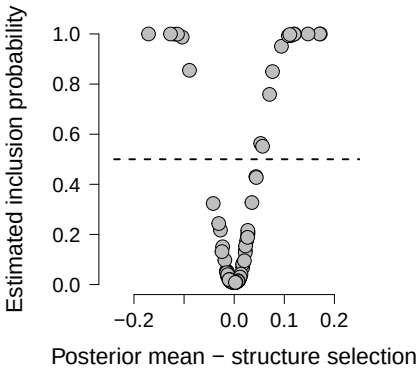
(c) Inclusion probabilities edge screening ($n = 2,000$).



(d) Inclusion probabilities edge screening ($n = 20,000$).



(e) Inclusion probabilities structure selection ($n = 2,000$).



(f) Inclusion probabilities structure selection ($n = 20,000$).

Figure 4. The top two panels show scatterplots of the posterior means and posterior modes of the association parameters that were obtained from our structure selection and edge screening procedures, respectively. The gray points are points of disagreement. The middle two panels show the edge screening inclusion probabilities and the bottom two panels the structure selection inclusion probabilities. The dashed lines are the bisection lines.

Network Analysis of Alcohol Use Disorder and Depression Data

For an empirical illustration of our Bayesian methods, we assess the relationship between symptoms of alcohol use disorder (AUD) and major depressive disorder (MDD) using data from the National Survey on Drug Use and Health (NSDUH; United States Department of Health and Human Services, 2016). The NSDUH is an American population study on tobacco, alcohol, and drug use, and mental health issues in the United States. The goal of the NSDUH is to provide accurate estimates on current patterns of substance abuse and its consequences for mental health. The survey is conducted in all 50 states, aiming at a sample of 70,000 individuals; participants have to be above the age of 12 and are randomly selected based on household addresses. We focus on the data on alcohol use and depression obtained in 2014.

The 2014 data comprises 55,271 participants. We exclude participants below the age of 18, that never drank alcohol, or that did not drink alcohol on more than six occasions in the past year. The final data analyzed here comprises 26,571 participants.

We included the seven items related to the DSM-V (American Psychiatric Association, 2013) criteria for AUD, and the nine symptoms in the NSDUH survey data comprising the DSM-V criteria for MDD in our analysis. The NSDUH derives the MDD symptoms from survey items formulated in a skip-structure. In this set-up, participants are allowed to skip certain items based on the answers they provide. Therefore, some specifics of symptoms are not assessed for participants, which may cause more absence scores for symptoms or problems than is the case.

In our analyses below, we first screen the network for promising edges, and then select plausible structures from the structure space instantiated by the set of promising edges. We will also perform structure selection without this initial pruning to illustrate the necessity of the edge screening step.

Edge Screening

In total, there were $p = 16$ variables, and $\binom{16}{2} = 120$ associations or possible edges to consider. We ran the edge screening procedure using EMVS on the selected NSDUH data. The EMVS set-up with a uniform prior on the structure space selected the same edges as the EMVS set-up with a uniform prior on structure complexity. We continue here using the results from the former. The edge screening procedure identified 62 promising edges, pruning almost half of the available connections. Edge screening using a uniform prior distribution on structure complexity gave the same results. The eLasso method identified 61 edges, three of which were not identified by our edge screening procedure. There were four edges identified by our edge screening procedure, that were not identified by eLasso. Figure 5a shows the network generated by the screened edges, where blue edges constitute positive associations, and red edges constitute negative associations.

We glean several important observations from Figure 5a. First, with 33 out of $\binom{9}{2} = 36$ possible connections between its nine symptoms, MDD appears to be densely connected. This result may be due, in part, to the skip structure that underlies the NSDUH assessment of MDD symptoms. However, it is in line with other results about MDD symptoms in the general population (e.g., Caspi et al., 2014). Second, with 20 out $\binom{7}{2} = 21$ possible connections between its seven symptoms, AUD also appears to be densely connected. The estimated associations are less strong than with MDD, which may be due to the skip structure that underlies the assessment of MDD symptoms. Third, there are relatively few estimated connections between the two disorders. Fourth, our edge screening

procedure identified a negative association between depressed mood and withdrawal symptoms. Negative associations are scarce in cross-sectional analyses, such as the one reported here.

Structure Selection

We identified 62 promising edges with our screening procedure, which generates a local median probability structure (LMS, c.f. Figure 5a). We now wish to find out what the plausible structures are for the data at hand and how the LMS in Figure 5a relates to the global median probability structure (GMS), i.e., the structure with edges that have marginal posterior inclusion probabilities

$$p(\gamma_{ij} = 1 \mid \mathbf{X}) = \sum_{\boldsymbol{\gamma}_s : \gamma_{ij}=1} p(\boldsymbol{\gamma}_s \mid \mathbf{X}) \geq \frac{1}{2}.$$

Barbieri and Berger (2004) showed that this GMS has, in general, excellent predictive properties. We again use the uniform prior on the structure space, which is consistent with the edge screening results shown above.

We ran the Gibbs sampler for 100,000 iterations, which visited 62 out of $2^{66} \approx 7e^{19}$ possible structures. Pitting the visited structures against the most frequently visited structure using the Bayes factor,[7]

$$\text{BF}_{1s} = \frac{p(\boldsymbol{\gamma}_1 \mid \mathbf{X})}{p(\boldsymbol{\gamma}_s \mid \mathbf{X})} = \frac{p^*(\mathbf{X} \mid \boldsymbol{\gamma}_1)}{p^*(\mathbf{X} \mid \boldsymbol{\gamma}_s)},$$

where $\boldsymbol{\gamma}_1$ denotes the most frequently visited model, we identified three structures for which the most visited structure was less than ten times as plausible. A Bayes factor BF_{1s} of ten or greater is often interpreted to provide strong evidence in favor of $\boldsymbol{\gamma}_1$ (see, for instance Jeffreys, 1961; Lee & Wagenmakers, 2013; Wagenmakers, Love, et al., 2018). The structures for which BF_{1s} was less than ten, and their estimated posterior structure probabilities, are shown in panels (b), (c) and (d) in Figure 5. The three structures only differed in the relations between the two disorders.

In Figure 6a, we plot the posterior inclusion probabilities obtained from the edge screening analysis against those obtained from the structure selection analysis on the pruned structure space. We glean two things from this figure. First, the local inclusion probabilities are at the extremes, i.e., the values zero and one, whereas the global inclusion probabilities show a broader range of values. This difference in separation was also observed in the analysis of simulated data in Figure 4. The bottom left corner comprises culled edges that have a zero probability of inclusion. Second, there is a great agreement about which edges are or are not in the median probability structure. The LMS and GMS differed in only one edge (indicated in white; points of agreement are in gray). In Figure 7, we plot the GMS and a difference plot, which reveals the differences between the LMS and GMS (red edges indicating edges that are in the LMS, but not the GMS). Figure 5e shows that the negative

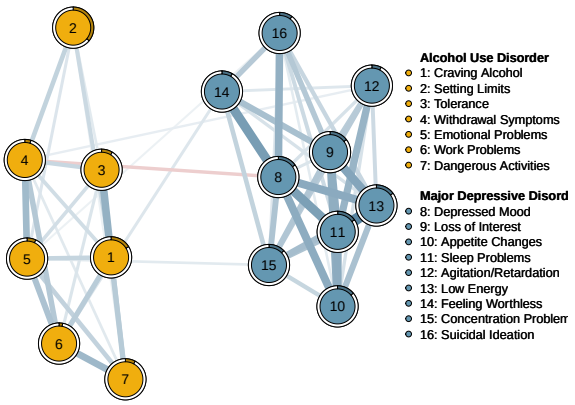
[7]For the hierarchical set-up, we have to include the prior structure probabilities in this equation:

$$\text{BF}_{1s} = \frac{p(\boldsymbol{\gamma}_1 \mid \mathbf{X})}{p(\boldsymbol{\gamma}_s \mid \mathbf{X})} \times \frac{p(\boldsymbol{\gamma}_s)}{p(\boldsymbol{\gamma}_1)} = \frac{p^*(\mathbf{X} \mid \boldsymbol{\gamma}_1)}{p^*(\mathbf{X} \mid \boldsymbol{\gamma}_s)}$$

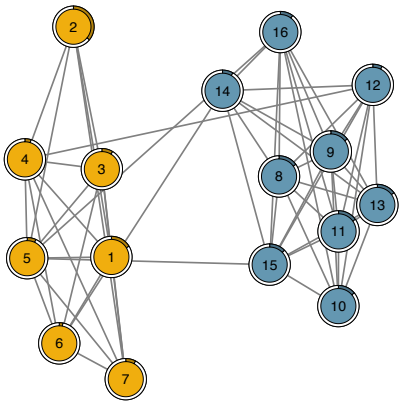
where $p(\boldsymbol{\gamma}_s)$ is the beta-binomial distribution

$$\binom{\binom{p}{2}}{\gamma_{s,++}} \frac{\text{B}\left(\gamma_{s,++} + \alpha, , \binom{p}{2} - \gamma_{s,++} + \beta\right)}{\text{B}\left(\alpha, , \beta\right)},$$

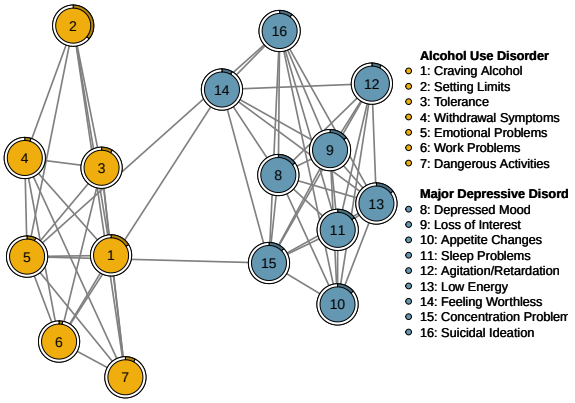
with $\alpha = \beta = 1$.



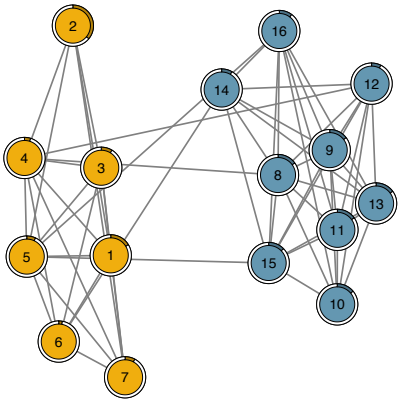
(a) Local median probability structure.



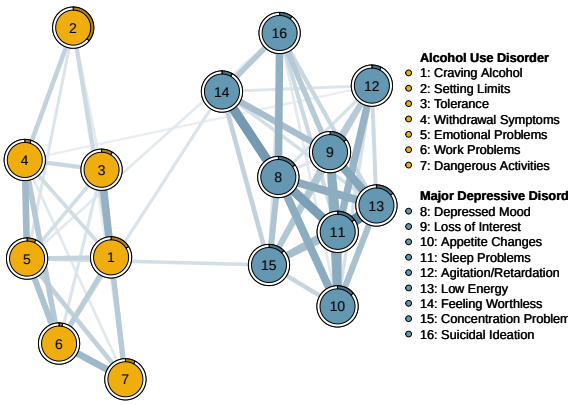
(b) First plausible structure: $p(\boldsymbol{\gamma}_1 \mid \mathbf{X}) = .607$.



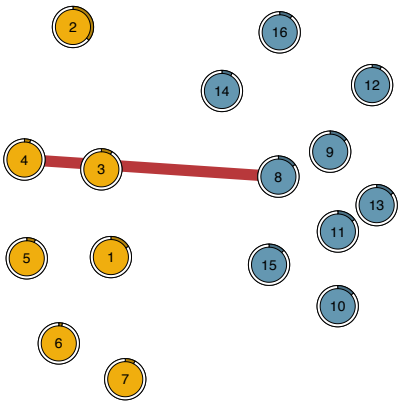
(c) Second plausible structure: $p(\boldsymbol{\gamma}_2 \mid \mathbf{X}) = .248$.



(d) Third plausible structure: $p(\boldsymbol{\gamma}_3 \mid \mathbf{X}) = .087$.

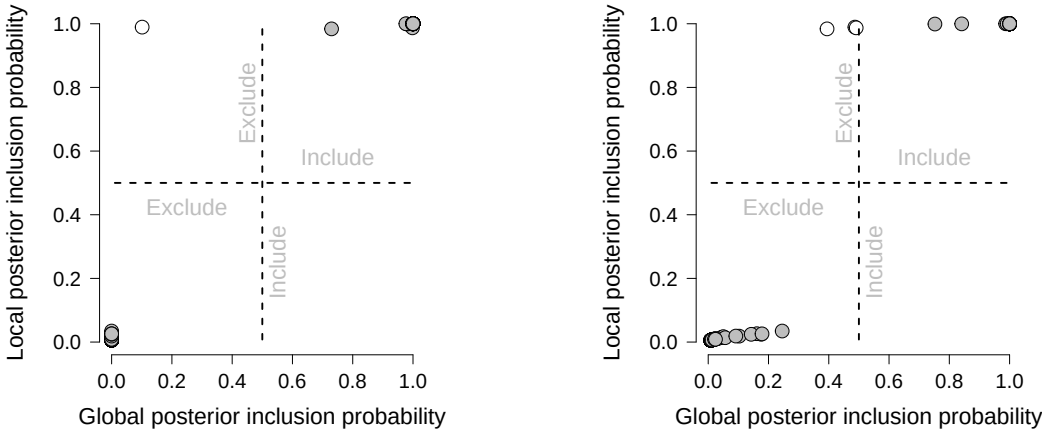


(e) Global median probability structure.



(f) Differences between median structures.

Figure 5. Edge screening and structure selection for NSDUH data. Panel (a) indicates the network generated by the promising edges identified by edge screening; the local median probability structure. Panels (b)-(d) indicate the three (most) plausible structures identified by structure selection on the pruned space. Panel (e) indicates the global median probability structure, and Panel (f) indicates the difference between the two median probability structures. The network plots are produced using the R package `qgraph` (Epskamp, Cramer, Waldorp, Schmittmann, & Borsboom, 2012).



(a) Pruned structure space

(b) Full structure space

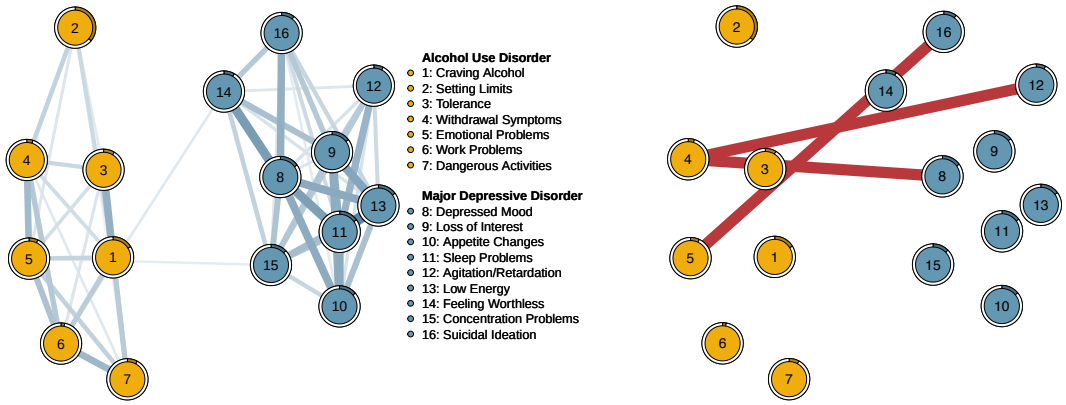
Figure 6. Plots of the local posterior inclusion probabilities of edges against the global posterior inclusion probabilities for the pruned space in Panel (a) and the full structure space in Panel (b). The dashed lines are the bisection lines.

association between nodes four and eight is not in the GMS ($p(\gamma_{4,8} = 1 \mid \mathbf{X}) = .101$). Thus, the LMS produced by our edge screening approach (c.f., Figure 5a) is an excellent approximation to the GMS identified on the pruned space (c.f., Figures 5e and 5f). Similar to our simulated example, the edge screening procedure proved to be more liberal than the structure selection approach, i.e., more edges were included in the LMS than in the GMS.

Structure Selection without Pruning

To analyze the benefit of our two-step procedure, with edge selection preceding structure selection to prune the structure space, we performed a structure selection analysis without pruning the structure space. We ran the Gibbs sampler for $100,000$ iterations, starting at the posterior mode, which visited $39,885$ out of $2^{120} \approx 1e^{36}$ possible structures. This result immediately underscores the importance of pruning the structure space before structure selection. The posterior structure probabilities of such a large collection of models cannot be estimated with great precision in a reasonable amount of time. Pitting the visited structures against the most frequently visited structure using the Bayes factor identified 52 plausible models. Two questions arise. The first question is about identifying the GMS and how it fares against the LMS identified with edge screening. Second, we wish to determine how the three previously identified structures stack up against the $39,885$ visited structures in the structure selection on the full structure space.

In Figure 6b, we plot the posterior inclusion probabilities obtained from the edge screening analysis against those obtained from the structure selection analysis on the full structure space. As before, the local inclusion probabilities are mostly located at the extreme ends of zero and one, whereas the global inclusion probabilities are more variable. This difference is emphasized in the bottom left corner of Figure 6b, since the previously culled edges now received non-zero probabilities. However, Figure 6b also reveals that there is great agreement about which edges are or are not in the median probability structure. The LMS and GMS on the full structure space differed



(a) Global median structure.

(b) Differences between median structures.

Figure 7. Application of Structure Selection to NSDUH data on the full structure space. Panel (a) indicates the global median probability structure on the full structure space, and Panel (b) indicates the difference between the local and global median probability structures. See text for details. The network plots are produced using the R package `qgraph` (Epskamp et al., 2012).

in three edges.

In Figure 7, we plot the GMS from the full space and a difference plot. Figure 7b shows that, as before with the pruned space, the negative association between nodes four and eight is not in the GMS ($p(\gamma_{4,8} = 1 \mid \mathbf{X}) = .487$). The edge between nodes four and twelve was also not in the second plausible structure observed before (c.f. Figure 5c; $p(\gamma_{4,12} = 1 \mid \mathbf{X}) = .394$). The edge between nodes five and sixteen, however, was not screened before ($p(\gamma_{5,16} = 1 \mid \mathbf{X}) = .491$). In sum, the LMS produced by our edge screening approach (c.f., Figure 5a) served as a good approximation to the GMS identified on the pruned space (c.f., Figures 5e and 5f) and on the full space (c.f., Figures 7a and 7b).

The Gibbs sampler on the full structure space visited $39,885$ structures. Of these $39,885$ structures, 75 were visited between 100 and $1,450$ times, indicating posterior probabilities between .0008 and .015. The remaining $39,785$ structures were visited less than 100 times, indicating a posterior probability of less than .0008. However, in total, the probabilities of these $39,785$ structures added up to .762. Thus, structure selection on the full structure space wastes valuable computational efforts on estimating insignificant structures. This is a prime example of dilution (George, 1999), and once more underscores the importance of pruning the structure space before performing structure selection. The posterior probabilities of the three structures identified earlier were .008, .015 and .008, and with that they were the 7^{th}, 1^{st} and 6^{th} most visited models, respectively. Nevertheless, given the vast amount of visited structures and the tiny probabilities associated with it, their estimates are highly uncertain.

Discussion

In this paper, we have introduced a novel objective spike-and-slab approach for structure selection for the Ising model, and we have illustrated the full suite of Bayesian tools using simulated

and empirical data. The empirical analysis allowed us to underscore the importance of trimming the structure space before its exploration, and that edge screening is capable of identifying relevant edges. The default specification of the spike-and-slab variances resulted in a selection method with consistently high specificity in our simulations, i.e., a low type-1 error rate in edge detection. In this respect, our procedure performs as good or even better than eLasso. Posterior estimates of the parameters are easy to obtain for both edge screening and structure selection procedures. Our structure selection procedure opened up the full spectrum of Bayesian tools, and, when paired with edge screening, it quickly zoomed in on plausible structures and promising effects. In sum, we have presented a complete Bayesian methodology for structure determination for the Ising model.

In a recent preprint, Bhattacharyya and Atchade (2019) also proposed a continuous spike-and-slab edge selection approach for the Ising model using the pseudolikelihood. The two methods were designed with a different focus, however. Whereas Bhattacharyya and Atchade focused on networks with many variables, we focused on psychological networks that are relatively small in comparison. As a result, the two approaches differ in several key aspects that make our approach more appealing to analyze psychological networks. For example, Bhattacharyya and Atchade did not trim the structure space before exploring it with a Gibbs sampler. Our empirical example illustrated why we believe that this is a bad idea. At the same time, we addressed some outstanding issues in this paper that Bhattacharyya and Atchade left open. For example, Bhattacharyya and Atchade (2019) analyzed the p full-conditionals in Eq. (2) in isolation, which provided them an opportunity for fast parallel processing. However, this also forced them to stipulate two independent prior distributions on each focal parameter, which means that they ended up with two posterior distributions for each association. Unfortunately, Bhattacharyya and Atchade provided no principled solution for combining these estimates for either structure selection or parameter estimation. Another issue is that their spike-and-slab approach required the specification of tuning parameters, but they offered no guidance or automated procedure for their specification. In sum, our method i) offers an objective specification of the prior distributions that lead to sensible answers, ii) trims the structure space to circumvent issues related to dilution, and iii) allows for a meaningful interpretation of the estimated posteriors. Despite these crucial differences, however, the approach of Bhattacharyya and Atchade is broader than ours, as they also analyzed networks of polytomous variables, while we exclusively focus on the binary case in this paper.

Our specification of the hyperparameters stipulates a mixture of two unit information priors, one fixed and one shrinking, that *a priori* match an approximate credible interval. We chose this set-up to mimic the eLasso approach of van Borkulo et al. (2014) and aimed for high specificity. However, researchers might have a different aim and wish to have methods available that have a higher sensitivity (e.g., see the considerations in Epskamp et al., 2017) or that aim for a low false discovery rate instead (e.g., Storey, 2003). In principle, penalty tuning procedures and prior structure probabilities could be tailored to achieve different goals. For example, we could adopt the eLasso approach and select the penalty ξ that minimizes the Bayesian information criterion (BIC; Schwarz, 1978) or the extended BIC (EBIC$_\lambda$, where λ is a penalty on complexity; Chen & Chen, 2008) instead of matching the spike-and-slab intersections to credible intervals. These two criteria usually achieve higher sensitivity than Lasso, and naturally tie in with the two prior distributions on the structure space that we have used here: A uniform prior distribution on the structure space is consistent with BIC, and a uniform prior distribution on structure complexity is compatible with EBIC$_1$. Furthermore, several alternative prior distributions that account for multiple testing have been discussed in the variable selection literature (e.g., Castillo, Schmidt-Hieber, & van der Vaart,

2015; Womack, Fuentes, & Taylor-Rodriguez, 2015). In sum, there are plenty of options to tailor the spike-and-slab approach to the specific needs of empirical researchers.

A caveat in our suite of Bayesian tools is the Bayes factor comparing two specific topologies. In principle, we can compute the Bayes factor from the posterior structure probabilities obtained from our structure selection procedure, but only if the Gibbs sampler visited the two structures under scrutiny. However, there is no guarantee that the Gibbs sampler visits the two structures, and even if the Gibbs sampler visits them, their estimated posterior probabilities can be uncertain. We need a more dedicated approach to estimate the Bayes factor comparing two particular structures of interest. We believe that the Laplace approximation that we have used in the paper will be a good starting point for this. However, the Laplace method's efficiency crucially depends on the accuracy of the normal approximation to the posterior distribution of parameters in the model. We have found that the normal approximation works very well in simulations. Another option would be the Bridge sampler (Gronau et al., 2017; Meng & Wong, 1996), which fits seamlessly with our Gibbs sampling approach.

Implementing our procedures in a compiled language is one of several improvements that we envision for our software. At this moment, our methods are wholly implemented in `R` (R Core Team, 2019). Our current implementation of the edge screening procedure implementation is a bit slower than the eLasso implementation in `IsingFit` —the analysis of NSDUH data took approximately 40 seconds for edge screening and 15 seconds for `IsingFit`— structure selection is considerably slower since the Gibbs sampler needs more time to explore the network space. There are currently are two computational bottlenecks: The specification of the Hessian matrix, and running the Gibbs sampler. Both involve iterating loops that can be computed much faster in a compiled language. Another aspect that we plan to implement shortly is the treatment of missing data. Two options present itself. The first uses selection functions for pairwise removal of missing data points; the second is data-augmentation or imputation. Both methods assume that data are missing at random, or are at least ignorable. The analysis of structurally missing data, e.g., missing data introduced by a skip structure as in our example, requires a different model set-up, in principle, and remains an open problem. As for different models, we are currently working on extending the method to Ising models for polytomous data (c.f., Bhattacharyya & Atchade, 2019). We also plan to implement our software in the open-source statistical software JASP (Love et al., 2019; Wagenmakers, Love, et al., 2018), which would build a user-friendly interface around our `R` package.

References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC.: American Psychiatric Association.

Arnold, B. C., & Strauss, D. (1991). Pseudolikelihood estimation: Some examples. *Sankhyā: The Indian Journal of Statistics, Series B.*, *53*(2), 233–243.

Barber, R. F., & Drton, M. (2015). High dimensional Ising model selection with Bayesian information criteria. *Electronic Journal of Statistics*, *9*(1), 567–607. doi: 10.1214/15-EJS1012

Barbieri, M. M., & Berger, J. O. (2004). Optimal predictive model selection. *Annals of Statistics*, *32*(3), 870–897. doi: 10.1214/009053604000000238

Besag, J. (1975). Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *24*(3), 179–195. doi: 10.2307/2987782

Bhattacharyya, A., & Atchade, Y. (2019). *Bayesian analysis of high-dimensional discrete graphical models.* (ArXiv 1907.01170)

Borsboom, D., & Cramer, A. O. J. (2013). Network analysis: An integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology*, *9*, 91–121. doi: 10.1146/annurev-clinpsy-050212-185608

Brusco, M., Steinley, D., Hoffman, M., Davis-Stober, C., & Wasserman, S. (in press). On Ising models and algorithms for the construction of symptom networks in psychopathological research. *Psychological Methods*, 1–19. doi: 10.1037/met0000207

Bühlmann, P., Kalisch, M., & Meier, L. (2014). High-dimensional statistics with a view toward applications in biology. *Annual Reviews of Statistics and Its Applications*, *1*, 255–278. doi: 10.1146/annurev-statistics-022513-115545

Caspi, A., Houts, R., Belsky, D., Goldman-Mellor, S., Harrington, H., Israel, S., ... Moffit, T. (2014). The p factor: One general psychopathology factor in the structure of psychiatric disorders? *Clinical Psychological Science*, *2*(2), 119–137. doi: 10.1177/2167702613497473

Castillo, I., Schmidt-Hieber, J., & van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics*, *43*(5), 1986–2018. doi: 10.1214/15-AOS1334

Chen, J., & Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, *95*(3), 759–771. doi: 10.1093/biomet/asn034

Consonni, G., Fouskakis, D., Liseo, B., & Ntzoufras, I. (2018). Prior distributions for objective Bayesian analysis. *Bayesian Analysis*, *13*(2), 627–679. doi: 10.1214/18-BA1103

Constantini, G., Richetin, J., Preti, E., Casini, E., Epskamp, S., & Perugi, M. (2019). Stability and variability of personality networks: A tutorial on recent developments in network psychometrics. *Personality and Individual Differences*, *136*, 68–78. doi: 10.1016/j.paid.2017.06.011

Cox, D. (1972). The analysis of multivariate binary data. *Journal of the Royal Statistical Society. Series B (Applied Statistics)*, *21*(2), 113–120. doi: 10.2307/2346482

Cramer, A. O. J., van Borkulo, C. D., Giltay, E. J., van der Maas, H. L. J., Kendler, K. S., Scheffer, M., & Borsboom, D. (2016). Major depression as a complex dynamic system. *PLoS One*, *11*(12), 1–20. doi: 10.1371/journal.pone.0167490

Cramer, A. O. J., van der Sluis, S., Noordhof, A., Wichers, M., Geschwind, N., Aggen, S. H., ... Borsboom, D. (2012). Dimensions of normal personality as networks in search of equilibrium: You can't like parties if you don't like people. *European Journal of Personality*, *26*, 414–431. doi: 10.1002/per.1866

Csiszár, I., & Talata, Z. (2006). Consistent estimation of the basic neighborhood of Markov random fields. *The Annals of Statistics*, *34*(1), 123–145. doi: 10.1214/009053605000000912

Dalege, J., Borsboom, D., van Harreveld, F., van den Berg, H., Conner, M., & van der Maas, H. L. J. (2016). Towards a formalized acount of attitudes: The causal attitude network (CAN) model. *Psychological Review*, *123*(1), 2–22. doi: 10.1037/a0039802

Dalege, J., Borsboom, D., van Harreveld, F., & van der Maas, H. L. J. (2019). A network perspective on political attitudes: Testing the connectivity hypothesis. *Social Psychological and Personality Science*, *10*(6), 746–756. doi: 10.1177/1948550618781062

Dellaportas, P., Forster, J. J., & Ntzoufras, I. (2002). On Bayesian model and variable selection using MCMC. *Statistics and Computing*, *12*, 27–36. doi: 10.1023/A:1013164120801199

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1–38. Retrieved from `https://www.jstor.org/stable/2984875`

Donner, C., & Opper, M. (2017). Inverse Ising problem in continuous time: A latent variable approach. *Physical Review E*, *96*(062104), 1–9. doi: 10.1103/PhysRevE.96.062104

Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods*, *50*, 195–212. doi: 10.3758/s13428-017-0862-1

Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). qgraph: Network visualizations of relationships in psychometric data. *Journal of Statistical Software*, *48*(4), 1–18. Retrieved from `http://www.jstatsoft.org/v48/i04/`

Epskamp, S., Kruis, J., & Marsman, M. (2017). Estimating psychopathological networks: Be careful what you wish for. *PLoS One*, *12*(e0179891). doi: 10.1371/journal.pone.0179891

Epskamp, S., Maris, G., Waldorp, L., & Borsboom, D. (2018). Network psychometrics. In P. Irwing, D. Hughes, & T. Booth (Eds.), *Handbook of psychometrics* (pp. 953–986). New York, NY: Wiley-Blackwell.

Erdős, P., & Rényi, A. (1960). On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, *5*(1), 17–60.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*(6), 721–741. doi: 10.1109/TPAMI.1984.4767596

George, E. I. (1999). Discussion of "Bayesian model averaging and model search strategies" by Clyde M. In J. Bernardo, J. Berger, A. Dawid, & A. Smith (Eds.), *Bayesian statistics* (Vol. 6, pp. 175–177). Oxford University Press.

George, E. I., & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, *88*(423), 881-889.

George, E. I., & McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, *7*(2), 339–373.

Geys, H., Molenberghs, G., & Ryan, L. M. (2007). Pseudo-likelihood inference for clustered binary data. *Communications in Statistics - Theory and Methods*, *26*(11), 2743–2767. doi: 10.1080/03610929708832075

Gronau, Q. F., Sarafoglou, A., Matzke, D., Boehm, U., Marsman, M., Leslie, D. S., … Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, *81*, 80–97. doi: 10.1016/j.jmp.2017.09.005

Ising, E. (1925). Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik*, *31*(1), 253–258. doi: 10.1007/BF02980577

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Oxford University Press.

Kass, R. E., & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relation to the Schwarz criterion. *Journal of the American Statistical Association*, *90*(431), 928–934. doi: 10.1080/01621459.1995.10476592

Kindermann, R., & Snell, J. L. (1980). *Markov random fields and their applications* (Vol. 1). Providence: American Mathematical Society.

Knight, K., & Fu, W. (2000). Asymptotics of Lasso-type estimators. *The Annals of Statistics*, *28*(5), 1356–1378. doi: 10.1214/aos/1015957397

Kuo, L., & Mallick, B. (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B.*, *60*(1), 65–81.

Kyung, M., Gill, J., Ghosh, M., & Casella, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*.

Lange, K. (1995). A gradient algorithm locally equivalent to the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological).*, *57*(2), 425–437. Retrieved from `https://www.jstor.org/stable/2345971`

Lee, M. D., & Wagenmakers, E. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, A. J., … Wagenmakers, E. (2019). JASP – graphical statistical software for common statistical designs. *Journal of Statistical Software*, *88*(2), 1–17. doi: 10.18637/jss.v088.i02

Makalic, E., & Schmidt, D. F. (2016). *High-dimensional Bayesian regularised regression with the BayesReg package.* (arXiv:1611.06649v3)

Marsman, M., Borsboom, D., Kruis, J., Epskamp, S., van Bork, R., Waldorp, L. J., … Maris, G. K. J. (2018). An introduction to network psychometrics: Relating Ising network models to item response theory models. *Multivariate Behavioral Research*, *53*(1), 15–35. doi: 10.1080/00273171.2017.1379379

Marsman, M., Maris, G. K. J., Bechger, T. M., & Glas, C. A. W. (2015). Bayesian inference for low-rank Ising networks. *Scientific Reports*, *5*(9050). doi: 10.1038/srep09050

Marsman, M., Tanis, C. C., Bechger, T. M., & Waldorp, L. J. (2019). Network psychometrics in educational

practice. Maximum likelihood estimation of the Curie-Weiss model. In B. P. Veldkamp & C. Sluijter (Eds.), *Theoretical and practical advances in computer-based educational measurement* (pp. 93–120). Cham, Switzerland: Springer.

Marsman, M., & Wagenmakers, E. (2017). Bayesian benefits with JASP. *European Journal of Developmental Psychology*, *14*(5), 545–555. doi: 10.1080/17405629.2016.1259614

Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*.

Meng, X., & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, *6*(4), 831–860.

Miller, J. W. (2019). Asymptotic normality, concentration, and coverage of generalized posteriors. *Retrieved from: https://arxiv.org/abs/1907.09611*. (ArXiv preprint)

Narisetty, N. N., & He, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*, *42*(2), 789–817. doi: 10.1214/14-AOS1207

Ntzoufras, I. (2009). *Bayesian modeling using WinBUGS*. Hoboken, N.J.: Wiley and Sons.

O'Hara, R. B., & Sillanpää, M. J. (2009). A review of Bayesian variable selection methods: What, how and which. *Bayesian Analysis*, *4*(1), 85–118. doi: 10.1214/09-BA403

Park, T., & Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, *103*(482), 681–686. doi: 10.1198/016214508000000337

Pensar, J., Nyman, H., Niiranen, J., & Corander, J. (2017). Marginal pseudo-likelihood learning of discrete Markov network structures. *Bayesian Analysis*, *12*(4), 1195–1215. doi: 10.1214/16-BA1032

Polson, N. G., Scott, J. G., & Windle, J. (2013a). Bayesian inference for logistic models using Pólya-Gamma latent variables. *Journal of the American Statistical Association*, *108*(504), 1339–1349. doi: 10.1080/01621459.2013.829001

Polson, N. G., Scott, J. G., & Windle, J. (2013b). *Bayesian inference for logistic models using Pólya-Gamma latent variables.* Retrieved from `http://arxiv.org/abs/1205.0310`

Pötscher, B. M., & Leeb, H. (2009). On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. *Journal of Multivariate Analysis*, *100*(9), 2065–2082. doi: 10.1016/j.jmva.2009.06.010

R Core Team. (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Raftery, A. E. (1999). Bayes factors and BIC. Comment on "A critique of the Bayesian information criterion for model selection". *Sociological Methods & Research*, *27*(3), 411–427. doi: 10.1177/0049124199027003005

Ravikumar, P., Wainwright, M. J., & Lafferty, J. D. (2010). High-dimensional Ising model selection using l_1-regularized logistic regression. *Annals of Statistics*, *38*(3), 1287–1319. doi: 10.1214/09-AOS691

Ročková, V. (2018). Bayesian estimation of sparse signals with a continuous spike-and-slab prior. *The Annals of Statistics*, *46*(1), 401–437. doi: 10.1214/17-AOS1554

Ročková, V., & George, E. I. (2014). EMVS: The EM approach to Bayesian variable selection. *Journal of the American Statistical Association*, *109*(506), 828–846. doi: 10.1080/01621459.2013.869223

Ročková, V., & George, E. I. (2018). The spike-and-slab lasso. *Journal of the American Statistical Association*, *113*(521), 431–444. doi: 10.1080/01621459.2016.1260469

Savi, A. O., Marsman, M., van der Maas, H. L. J., & Maris, G. K. J. (2019). The wiring of intelligence. *Perspectives on Psychological Science*, *16*(6), 1034–1061. doi: 10.1177/1745691619866447

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. doi: 10.1214/aos/1176344136

Scott, J. G., & Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Annals of Statistics*, *38*(5), 2587–2619. doi: 10.1214/10-AOS792

Storey, J. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *The Annals of Statistics*, *31*(6), 2013–2035. doi: 10.1214/aos/1074290335

Tanner, M. (1996). *Tools for statistical inference. methods for the exploration of posterior distributions and likelihood functions.* New York, NY: Springer-Verlag. doi: 10.1007/978-1-4612-4024-2

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288. Retrieved from https://www.jstor.org/stable/2346178

Tierney, L., Kass, R. E., & Kadane, J. B. (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *Journal of the American Statistical Association*, *84*(407), 710–716. doi: 10.2307/2289652drton

United States Department of Health and Human Services. (2016). *National survey on drug use and health, 2014.* Inter-university Consortium for Political and Social Research [distributor]. doi: 10.3886/ICPSR36361.v1

van Borkulo, C. D., Borsboom, D., Epskamp, S., Blanken, T. F., Boschloo, L., Schoevers, R. A., & Waldorp, L. J. (2014). A new method for constructing networks from binary data. *Scientific Reports*, *4*((5918)). doi: 10.1038/srep05918

van Borkulo, C. D., Epskamp, S., & Robitzsch, A. (2016). IsingFit: Fitting Ising models using the eLasso method [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=IsingFit (R package version 0.3.1)

van de Geer, S., Bülmann, P., Ritov, Y., & Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, *42*(3), 1166–1202. doi: 10.1214/14-AOS1221

van der Maas, H. L. J., Kan, K.-J., Marsman, M., & Stevenson, C. E. (2017). Network models for cognitive development and intelligence. *Journal of Intelligence*, *5*(2), 1–17. doi: 10.3390/jintelligence5020016

Wagenmakers, E. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*, 779–804. doi: 10.3758/BF03194105

Wagenmakers, E., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., … Morey, R. D. (2018). Bayesian inference for psychology. Part ii: Example applications with JASP. *Psychonomic Bulletin & Review*, *25*(1), 58–76. doi: 10.3758/s13423-017-1323-7

Wagenmakers, E., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., … Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulleting & Review*, *25*(1), 58–76. doi: 10.3758/s13423-017-1343-3

Windle, J., Polson, N. G., & Scott, J. G. (2014). *Sampling Pólya-Gamma random variates: Alternate and approximate techniques.* (ArXiv: 1405.0506)

Womack, A. J., Fuentes, C., & Taylor-Rodriguez, D. (2015). *Model space priors for objective sparse Bayesian regression.* (ArXiv: 1511.04745)

Appendix A
EM Variable Selection for the Pseudolikelihood Ising Model

The E-Step. The Q-function factors into three distinct terms[8]:

$$\begin{aligned} &\mathrm{Q}\left(\boldsymbol{\Sigma}, \boldsymbol{\mu}, \theta \mid \boldsymbol{\Sigma}^k, \theta^k\right) \\ &\quad = \mathrm{Q}_1(\boldsymbol{\Sigma}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma}^k, \theta^k) + \mathrm{Q}_2(\theta \mid \boldsymbol{\Sigma}^k, \theta^k) + \mathrm{C}, \end{aligned} \tag{9}$$

where $\mathrm{C} = -\ln(p(\mathbf{X}))$ is a constant term.

The first term in Eq. (9) concerns the pseudoposterior of the Ising model's parameters

$$\begin{aligned} \mathrm{Q}_1(\boldsymbol{\Sigma}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma}^k, \theta^k) = \ln\left(p^*(\mathbf{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})\right) + \ln\left(p(\boldsymbol{\mu})\right) \\ + \mathbb{E}\left(\ln\left(p(\boldsymbol{\Sigma} \mid \boldsymbol{\gamma})\right) \mid \boldsymbol{\Sigma}^k, \theta^k\right), \end{aligned}$$

[8] Assuming that θ is assigned a prior distribution. Otherwise the second term is omitted.

and involves the expectation of the log-transformed spike-and-slab prior

$$\mathbb{E}\Big(\ln\left(p(\boldsymbol{\Sigma} \mid \boldsymbol{\gamma})\right) \mid \boldsymbol{\Sigma}^k, \theta^k\Big) = \\ \mathrm{C}_1 - \frac{1}{2}\sum_{i=1}^{p-1}\sum_{j=i+1}^{p} \sigma_{ij}^2\, \mathbb{E}\left(\frac{1}{\gamma_{ij}\nu_1 + (1-\gamma_{ij})\nu_0}\,\middle|\, \sigma_{ij}^k, \theta^k\right),$$

where C_1 is a constant term, and the last term can be reformulated as (c.f., Ročková & George, 2014, Eq. 3.6)

$$-\frac{1}{2}\sum_{i=1}^{p-1}\sum_{j=i+1}^{p} \sigma_{ij}^2 \left\{\frac{\mathbb{E}\left(\gamma_{ij}\,\middle|\, \sigma_{ij}^k, \theta^k\right)}{\nu_1} + \frac{1-\mathbb{E}\left(\gamma_{ij}\,\middle|\, \sigma_{ij}^k, \theta^k\right)}{\nu_0}\right\},$$

where the posterior expectation of the selection variable is equal to

$$\mathbb{E}\left(\gamma_{ij}\,\middle|\, \sigma_{ij}^k, \theta^k\right) = \frac{p(\sigma_{ij}^k \mid \gamma_{ij}=1)\, p(\gamma_{ij}=1 \mid \theta^k)}{\sum_{\Gamma_{ij}=\gamma_{ij}}\, p(\sigma_{ij}^k \mid \gamma_{ij})\, p(\gamma_{ij} \mid \theta^k)}. \tag{10}$$

The second term in Eq. (9) concerns the posterior distribution of the prior inclusion probability

$$\mathrm{Q}_2(\theta \mid \boldsymbol{\Sigma}^k, \theta^k) = \mathbb{E}\Big(\ln\left(p(\boldsymbol{\gamma} \mid \theta)\right) \mid \boldsymbol{\Sigma}^k, \theta^k\Big) + \ln\left(p(\theta)\right),$$

and involves the expectation of the log-transformed prior distribution on the selector variables

$$\mathbb{E}\Big(\ln\left(p(\boldsymbol{\gamma} \mid \theta)\right) \mid \boldsymbol{\Sigma}^k, \theta^k\Big) = \\ \ln(\theta)\binom{p}{2} + \ln\left(\frac{\theta}{1-\theta}\right)\sum_{i=1}^{p-1}\sum_{j=i+1}^{p} \mathbb{E}\left(\gamma_{ij}\,\middle|\, \sigma_{ij}^k, \theta^k\right),$$

and is also readily computed using the expression in Eq. (10).

The M-Step. We separately optimize the two components of the Q-function in the M-step. Unfortunately, there is no closed-form solution for the maximization of Q_1, and we approximate the M-Step using a single iteration of a Newton-Raphson algorithm (Lange, 1995; Tanner, 1996). The details are in Appendix B. The maximization of Q_2 is in closed-form,

$$\theta^{k+1} = \frac{\sum_{i=1}^{p-1}\sum_{j=i+1}^{p} \mathbb{E}\left(\gamma_{ij} \mid \sigma_{ij}^k, \theta^{(k)}\right) + \alpha - 1}{\alpha + \beta + \binom{p}{2} - 2}.$$

Posterior Standard Deviations. The EM algorithm provides us with an estimate of a local posterior mode, and we seek a way to quantify the uncertainty in this modal estimate. We express this uncertainty using the variance-covariance matrix of the normal approximation to the posterior (Tanner, 1996, e.g.,), i.e., the inverse of the Hessian matrix. The Hessian matrix is computed in the M-step of our EMVS approach, see Appendix B, and serves as an estimate of the variance-covariance matrix of the complete posterior $p(\boldsymbol{\Sigma}, \boldsymbol{\mu}, \theta, \boldsymbol{\gamma} \mid \mathbf{X})$. To estimate the variance-covariance

matrix of the marginal posterior $p(\boldsymbol{\Sigma}, \boldsymbol{\mu}, \theta \mid \mathbf{X})$ we have to use the inverse of the Hessian subject to the marginal spike-and-slab prior distributions on the interaction effects,

$$p(\sigma_{ij} \mid \theta) = \theta \frac{1}{\sqrt{2\pi\nu_1}} e^{\left(-\frac{1}{2\nu_1}\sigma_{ij}^2\right)} + (1-\theta) \frac{1}{\sqrt{2\pi\nu_0}} e^{\left(-\frac{1}{2\nu_0}\sigma_{ij}^2\right)}.$$

For prior specification —setting the values of ν_1— we use the inverse Hessian excluding prior distributions on the parameters.

Appendix B
The M-Step approximation for Q_1

We approximate the M-step of Q_1 using a single iteration of the Newton-Raphson algorithm. Let $\boldsymbol{\eta} = \{\mu_1, \ldots, \mu_p, \sigma_{12}, \ldots, \sigma_{(p-1)p}\}$ denote the $\left(\binom{p}{2} + p\right) \times 1$ vector of pseudolikelihood parameters. Then, the Newton-Raphson iteration is equal to

$$\boldsymbol{\eta}^{k+1} = \boldsymbol{\eta}^k + \mathbf{H}^{-1}\mathbf{D},$$

where $\mathbf{D}$ is the $\left(\binom{p}{2} + p\right) \times 1$ vector of first-order partial derivatives and $\mathbf{H}$ the $\left(\binom{p}{2} + p\right) \times \left(\binom{p}{2} + p\right)$ matrix of second-order partial derivatives, i.e., the Hessian matrix.

The first-order partial derivatives are equal to

$$\frac{\partial}{\partial \mu_i} Q_1 = \sum_{v=1}^{n} x_{vi} - \sum_{v=1}^{n} \frac{\exp\left(\mu_i + \sum_{j\neq i} \sigma_{ij} x_{vj}\right)}{1 + \exp\left(\mu_i + \sum_{j\neq i} \sigma_{ij} x_{vj}\right)} - \mu_i$$

for the main effects, and

$$\begin{aligned}
\frac{\partial}{\partial \sigma_{ij}} Q_1 &= 2\sum_{v=1}^{n} x_{vi}x_{vj} \\
&\quad - \sum_{v=1}^{n} x_{vj} \frac{\exp\left(\mu_i + \sum_{q\neq i} \sigma_{iq} x_{vq}\right)}{1 + \exp\left(\mu_i + \sum_{q\neq i} \sigma_{iq} x_{vq}\right)} \\
&\quad - \sum_{v=1}^{n} x_{vi} \frac{\exp\left(\mu_j + \sum_{q\neq j} \sigma_{jq} x_{vq}\right)}{1 + \exp\left(\mu_j + \sum_{q\neq j} \sigma_{jq} x_{vq}\right)} \\
&\quad - \sigma_{ij} \left\{ \frac{\mathbb{E}(\gamma_{ij} \mid \sigma_{ij}^k, \theta^k)}{\nu_1} + \frac{1 - \mathbb{E}(\gamma_{ij} \mid \sigma_{ij}^k, \theta^k)}{\nu_0} \right\} \\
&= 2\sum_{v=1}^{n} x_{vi}x_{vj} - \sum_{v=1}^{n} x_{vj}\, p_{vi}^* - \sum_{v=1}^{n} x_{vi}\, p_{vj}^* - \sigma_{ij} e_{ij}
\end{aligned}$$

for the interaction effects. Here, we have used p_{vi}^* to denote the conditional probability $p(X_i = 1 \mid \mathbf{x}_v^{(i)})$ and e_{ij} to denote the expected precision.

The Hessian matrix is slightly more complicated, as it requires some tedious bookkeeping. To emphasize its structure and ease its derivation we split the Hessian matrix in four components,

$$\mathbf{H} = \begin{pmatrix} \mathbf{H}_{\mu} & \mathbf{H}_{\mu\Sigma} \\ \mathbf{H}_{\mu\Sigma}^{\mathsf{T}} & \mathbf{H}_{\Sigma} \end{pmatrix},$$

where $\mathbf{H}_\mu$, and $\mathbf{H}_\Sigma$ are the second-order partial derivatives of the main effects $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively, and $\mathbf{H}_{\mu\Sigma}$ their cross-derivatives. The submatrix $\mathbf{H}_\mu$ is diagonal and has elements

$$\frac{\partial^2}{\partial\mu_i\partial\mu_j}\mathrm{Q}_1 = \begin{cases} -\sum_{v=1}^{n} p^*_{vi}(1-p^*_{vi}) - 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

The submatrix $\mathbf{H}_{\mu\Sigma}$ has elements

$$\frac{\partial^2}{\partial\mu_i\partial\sigma_{rj}}\mathrm{Q}_1 = \begin{cases} -\sum_{v=1}^{n} x_{vj}\, p^*_{vi}(1-p^*_{vi}) & \text{if } r = i \\ 0 & \text{if } i \neq r \end{cases}$$

Finally, the submatrix $\mathbf{H}_\Sigma$ has diagonal elements

$$\frac{\partial^2}{\partial\sigma^2_{ij}}\mathrm{Q}_1 = -\sum_{v=1}^{n} x_{vj}\, p^*_{vi}q^*_{vi} - \sum_{v=1}^{n} x_{vi}\, p^*_{vj}q^*_{vj} - e_{ij},$$

and off-diagonal elements

$$\frac{\partial^2}{\partial\sigma_{ij}\partial\sigma_{ru}}\mathrm{Q}_1 = \begin{cases} -\sum_{v=1}^{n} x_{vj}x_{vu}\, p^*_{vi}q^*_{vi} & \text{if } i = r \text{ and } j \neq u \\ -\sum_{v=1}^{n} x_{vj}x_{vr}\, p^*_{vi}q^*_{vi} & \text{if } i = u \text{ and } j \neq r \\ -\sum_{v=1}^{n} x_{vi}x_{vr}\, p^*_{vj}q^*_{vj} & \text{if } i \neq r \text{ and } j = u \\ -\sum_{v=1}^{n} x_{vi}x_{vu}\, p^*_{vj}q^*_{vj} & \text{if } i \neq u \text{ and } j = r \\ 0 & \text{if } i \neq u \text{ and } j \neq r \end{cases}$$

where $q^*_{vj} = 1 - p^*_{vj}$.

Appendix C
A Gibbs Sampling Routine for Structure Selection

The Gibbs sampler iterates between the following five steps. If a uniform prior is stipulated on the structure space, then step four is skipped.

Step 1. Sampling the main effects μ_i. With the assumption of prior independence, the main effects are also found to be independent *a posteriori*, and do not depend on the selection variables $\boldsymbol{\gamma}$. Given the standard normal prior distribution, the full-conditional posterior distribution $p(\mu_i \mid \boldsymbol{\sigma}_i, \boldsymbol{\omega}_i, \mathbf{X})$ of the main effect μ_i is a normal distribution, with mean

$$\frac{x_{i+} - \frac{n}{2} - \sum_{v=1}^{n}\sum_{j\neq i}\sigma_{ij}x_{jv}\omega_{iv}}{1+\omega_{i+}}$$

and variance $(1+\omega_{i+})^{-1}$, where we have used $\boldsymbol{\sigma}$ to denote the $p-1\times 1$ vector

$$\boldsymbol{\sigma} = (\sigma_{i1},\ \ldots,\ \sigma_{i(i-1))},\ \sigma_{i(i+1)},\ \ldots,\ \sigma_{ip})^{\mathsf{T}},$$

and x_{i+} and ω_{+i} to denote the margins $\sum_{v=1}^{n} x_{iv}$ and $\sum_{v=1}^{n}\omega_{vi}$, respectively.

Step 2. Sampling the interaction effects σ_{ij}. Given γ_{ij}, the prior distribution for σ_{ij} is normal with a zero mean, and variance $\phi = \gamma_{ij}\nu_1 + (1-\gamma_{ij})\nu_0$. The full-conditional posterior distribution is then a normal distribution with mean

$$\left(\boldsymbol{\omega}_i^\intercal \mathbf{x}_j + \boldsymbol{\omega}_j^\intercal \mathbf{x}_i + \phi^{-1}\right)^{-1} \times \left(2\mathbf{x}_i^\intercal \mathbf{x}_j - \frac{1}{2}x_{+i} - \frac{1}{2}x_{+j} - \sum_{v=1}^{n} \omega_{vi}x_{vj}\left[\mu_i + \sum_{q\neq i\neq j} \sigma_{iq}x_{vq}\right] - \sum_{v=1}^{n} \omega_{vj}x_{vi}\left[\mu_j + \sum_{q\neq i\neq j} \sigma_{jq}x_{vq}\right]\right)$$

and variance $\left(\boldsymbol{\omega}_i^\intercal \mathbf{x}_j + \boldsymbol{\omega}_j^\intercal \mathbf{x}_i + \phi^{-1}\right)^{-1}$, where we have used $\mathbf{x}_i$ to denote the $n \times 1$ vector with elements $[x_{iv}]$.

Step 3. Sampling the inclusion variables γ_{ij}. The full-conditional posterior distribution of γ_{ij} is a Bernoulli distribution with probability of inclusion:

$$p(\gamma_{ij} = 1 \mid \sigma_{ij}, \theta) = \frac{1}{1 + \frac{1-\theta}{\theta}\,\exp\left(\frac{1}{2(\nu_1-\nu_0)}\sigma_{ij}^2\right)}.$$

Step 4. Sampling the prior inclusion probability θ. The full-conditional posterior distribution of θ is a Beta distribution, with parameters

$$\alpha = 1 + \gamma_{++}/2,$$
$$\beta = 1 + \binom{p}{2} - \gamma_{++}/2,$$

where $\gamma_{++} = \sum_{i=1}^{p}\sum_{j=1}^{p}\gamma_{ij}$.

Step 5. Sampling the augmented variables ω_{vi}. The full-conditional posterior distribution of ω_{vi} is proportional to

$$p\left(\omega_{vi} \mid \boldsymbol{\sigma}_i, \mathbf{x}_v^{(i)}\right) \propto \exp\left(-\frac{1}{2}\omega_{vi}\left(\mu_i + \sum_{j\neq i}\sigma_{ij}x_{vj}\right)^2\right) p(\omega_{vi}),$$

where $p(\omega_{vi}) = p(\omega_{vi} \mid 1, 0)$ is a Pólya-Gamma distribution with parameters $b = 1$ and $c = 0$. Polson et al. (2013a) show that the Pólya-Gamma distribution with parameters $b = 1$ and $c \neq 1$ is equal to an exponential tilting of the Pólya-Gamma distribution with parameters $b = 1$ and $c = 0$,

$$p(\omega \mid 1, c) = \frac{\exp\left(-\frac{1}{2}c^2\omega\right)\, p(\omega \mid 1, 0)}{\cosh\left(\frac{1}{2}c\right)},$$

which consequently shows that $p\left(\omega_{vi} \mid \boldsymbol{\sigma}_i, \mathbf{x}_v^{(j)}\right)$ is a Pólya-Gamma distribution with parameters $b = 1$ and $c = \mu_i + \sum_{j\neq i}\sigma_{ij}x_{vj}$. These values can be simulated using the `R` (R Core Team, 2019) programs `BayesLogit` (Polson, Scott, & Windle, 2013b) and `BayesReg` (Makalic & Schmidt, 2016).