Extended Bayesian Information Criteria for Gaussian Graphical Models

Rina Foygel
University of Chicago
rina@uchicago.edu

Mathias Drton
University of Chicago
drton@uchicago.edu

Abstract

Gaussian graphical models with sparsity in the inverse covariance matrix are of significant interest in many modern applications. For the problem of recovering the graphical structure, information criteria provide useful optimization objectives for algorithms searching through sets of graphs or for selection of tuning parameters of other methods such as the graphical lasso, which is a likelihood penalization technique. In this paper we establish the consistency of an extended Bayesian information criterion for Gaussian graphical models in a scenario where both the number of variables p and the sample size n grow. Compared to earlier work on the regression case, our treatment allows for growth in the number of non-zero parameters in the true model, which is necessary in order to cover connected graphs. We demonstrate the performance of this criterion on simulated data when used in conjunction with the graphical lasso, and verify that the criterion indeed performs better than either cross-validation or the ordinary Bayesian information criterion when p and the number of non-zero parameters q both scale with n.

1 Introduction

This paper is concerned with the problem of model selection (or structure learning) in Gaussian graphical modelling. A Gaussian graphical model for a random vector $X = (X_1, \ldots, X_p)$ is determined by a graph G on p nodes. The model comprises all multivariate normal distributions $N(\mu, \Theta^{-1})$ whose inverse covariance matrix satisfies that $\Theta_{jk} = 0$ when $\{j, k\}$ is not an edge in G. For background on these models, including a discussion of the conditional independence interpretation of the graph, we refer the reader to [1].

In many applications, in particular in the analysis of gene expression data, inference of the graph G is of significant interest. Information criteria provide an important tool for this problem. They provide the objective to be minimized in (heuristic) searches over the space of graphs and are sometimes used to select tuning parameters in other methods such as the graphical lasso of [2]. In this work we study an extended Bayesian information criterion (BIC) for Gaussian graphical models. Given a sample of n independent and identically distributed observations, this criterion takes the form

$$BIC_\gamma(\mathbf{E}) = -2l_n(\hat{\Theta}(\mathbf{E})) + |\mathbf{E}| \log n + 4|\mathbf{E}|\gamma \log p, \tag{1}$$

where $\mathbf{E}$ is the edge set of a candidate graph and $l_n(\hat{\Theta}(\mathbf{E}))$ denotes the maximized log-likelihood function of the associated model. (In this context an edge set comprises unordered pairs $\{j, k\}$ of distinct elements in $\{1, \ldots, p\}$.) The criterion is indexed by a parameter $\gamma \in [0, 1]$; see the Bayesian interpretation of γ given in [3]. If $\gamma = 0$, then the classical BIC of [4] is recovered, which is well known to lead to (asymptotically) consistent model selection in the setting of fixed number of variables p and growing sample size n. Consistency is understood to mean selection of the smallest true graph whose edge set we denote $\mathbf{E}_0$. Positive γ leads to stronger penalization of large graphs and our main result states that the (asymptotic) consistency of an exhaustive search over a restricted

model space may then also hold in a scenario where p grows moderately with n (see the Main Theorem in Section 2). Our numerical work demonstrates that positive values of γ indeed lead to improved graph inference when p and n are of comparable size (Section 3).

The choice of the criterion in (1) is in analogy to a similar criterion for regression models that was first proposed in [5] and theoretically studied in [3, 6]. Our theoretical study employs ideas from these latter two papers as well as distribution theory available for decomposable graphical models. As mentioned above, we treat an exhaustive search over a restricted model space that contains all decomposable models given by an edge set of cardinality $|\mathbf{E}| \leq q$. One difference to the regression treatment of [3, 6] is that we do not fix the dimension bound q nor the dimension $|\mathbf{E}_0|$ of the smallest true model. This is necessary for connected graphs to be covered by our work.

In practice, an exhaustive search is infeasible even for moderate values of p and q. Therefore, we must choose some method for preselecting a smaller set of models, each of which is then scored by applying the extended BIC (EBIC). Our simulations show that the combination of EBIC and graphical lasso gives good results well beyond the realm of the assumptions made in our theoretical analysis. This combination is consistent in settings where both the lasso and the exhaustive search are consistent but in light of the good theoretical properties of lasso procedures (see [7]), studying this particular combination in itself would be an interesting topic for future work.

2 Consistency of the extended BIC for Gaussian graphical models

2.1 Notation and definitions

In the sequel we make no distinction between the edge set $\mathbf{E}$ of a graph on p nodes and the associated Gaussian graphical model. Without loss of generality we assume a zero mean vector for all distributions in the model. We also refer to $\mathbf{E}$ as a set of entries in a $p \times p$ matrix, meaning the $2|\mathbf{E}|$ entries indexed by (j,k) and (k,j) for each $\{j,k\} \in \mathbf{E}$. We use Δ to denote the index pairs (j,j) for the diagonal entries of the matrix.

Let Θ_0 be a positive definite matrix supported on $\Delta \cup \mathbf{E}_0$. In other words, the non-zero entries of Θ_0 are precisely the diagonal entries as well as the off-diagonal positions indexed by $\mathbf{E}_0$; note that a single edge in $\mathbf{E}_0$ corresponds to two positions in the matrix due to symmetry. Suppose the random vectors $X_1, \ldots, X_n$ are independent and distributed identically according to $N(0, \Theta_0^{-1})$. Let $S = \frac{1}{n}\sum_i X_i X_i^T$ be the sample covariance matrix. The Gaussian log-likelihood function simplifies to

$$l_n(\Theta) = \frac{n}{2}\left[\log\det(\Theta) - \operatorname{trace}(S\Theta)\right]. \tag{2}$$

We introduce some further notation. First, we define the maximum variance of the individual nodes:

$$\sigma^2_{\max} = \max_j (\Theta_0^{-1})_{jj}.$$

Next, we define $\theta_0 = \min_{\mathbf{e}\in\mathbf{E}_0} |(\Theta_0)_\mathbf{e}|$, the minimum signal over the edges present in the graph. (For edge $\mathbf{e} = \{j,k\}$, let $(\Theta_0)_\mathbf{e} = (\Theta_0)_{jk} = (\Theta_0)_{kj}$.) Finally, we write $\lambda_{\max}$ for the maximum eigenvalue of Θ_0. Observe that the product $\sigma^2_{\max}\lambda_{\max}$ is no larger than the condition number of Θ_0 because $1/\lambda_{\min}(\Theta_0) = \lambda_{\max}(\Theta_0^{-1}) \geq \sigma^2_{\max}$.

2.2 Main result

Suppose that n tends to infinity with the following asymptotic assumptions on data and model:

$$\begin{cases} \mathbf{E}_0 \text{ is decomposable, with } |\mathbf{E}_0| \leq q, \\ \sigma^2_{\max}\lambda_{\max} \leq C, \\ p = \mathbf{O}(n^\kappa),\ p \to \infty, \\ \gamma_0 = \gamma - (1 - \frac{1}{4\kappa}) > 0, \\ (p+2q)\log p \times \frac{\lambda^2_{\max}}{\theta_0^2} = \mathbf{o}(n) \end{cases} \tag{3}$$

Here $C, \kappa > 0$ and γ are fixed reals, while the integers p, q, the edge set $\mathbf{E}_0$, the matrix Θ_0, and thus the quantities $\sigma^2_{\max}$, $\lambda_{\max}$ and θ_0 are implicitly allowed to vary with n. We suppress this latter dependence on n in the notation. The 'big oh' $\mathbf{O}(\cdot)$ and the 'small oh' $\mathbf{o}(\cdot)$ are the Landau symbols.

Main Theorem. *Suppose that conditions (3) hold. Let $\mathcal{E}$ be the set of all decomposable models $\mathbf{E}$ with $|\mathbf{E}| \leq q$. Then with probability tending to 1 as $n \to \infty$,*

$$\mathbf{E}_0 = \arg\min_{\mathbf{E}\in\mathcal{E}} \mathrm{BIC}_\gamma(\mathbf{E}).$$

That is, the extended BIC with parameter γ selects the smallest true model $\mathbf{E}_0$ when applied to any subset of $\mathcal{E}$ containing $\mathbf{E}_0$.

In order to prove this theorem we use two techniques for comparing likelihoods of different models. Firstly, in Chen and Chen's work on the GLM case [6], the Taylor approximation to the log-likelihood function is used and we will proceed similarly when comparing the smallest true model $\mathbf{E}_0$ to models $\mathbf{E}$ which do not contain $\mathbf{E}_0$. The technique produces a lower bound on the decrease in likelihood when the true model is replaced by a false model.

Theorem 1. *Suppose that conditions (3) hold. Let $\mathcal{E}_1$ be the set of models $\mathbf{E}$ with $\mathbf{E} \not\supset \mathbf{E}_0$ and $|\mathbf{E}| \leq q$. Then with probability tending to 1 as $n \to \infty$,*

$$l_n(\Theta_0) - l_n(\hat{\Theta}(\mathbf{E})) > 2q(\log p)(1+\gamma_0) \quad \forall\, \mathbf{E} \in \mathcal{E}_1.$$

Secondly, Porteous [8] shows that in the case of two nested models which are both decomposable, the likelihood ratio (at the maximum likelihood estimates) follows a distribution that can be expressed exactly as a log product of Beta distributions. We will use this to address the comparison between the model $\mathbf{E}_0$ and decomposable models $\mathbf{E}$ containing $\mathbf{E}_0$ and obtain an upper bound on the improvement in likelihood when the true model is expanded to a larger decomposable model.

Theorem 2. *Suppose that conditions (3) hold. Let $\mathcal{E}_0$ be the set of decomposable models $\mathbf{E}$ with $\mathbf{E} \supset \mathbf{E}_0$ and $|\mathbf{E}| \leq q$. Then with probability tending to 1 as $n \to \infty$,*

$$l_n(\hat{\Theta}(\mathbf{E})) - l_n(\hat{\Theta}(\mathbf{E}_0)) < 2(1+\gamma_0)(|\mathbf{E}| - |\mathbf{E}_0|)\log p \quad \forall \mathbf{E} \in \mathcal{E}_0 \backslash \{\mathbf{E}_0\}.$$

Proof of the Main Theorem. With probability tending to 1 as $n \to \infty$, both of the conclusions of Theorems 1 and 2 hold. We will show that both conclusions holding simultaneously implies the desired result.

Observe that $\mathcal{E} \subset \mathcal{E}_0 \cup \mathcal{E}_1$. Choose any $\mathbf{E} \in \mathcal{E}\backslash\{\mathbf{E}_0\}$. If $\mathbf{E} \in \mathcal{E}_0$, then (by Theorem 2):

$$\mathrm{BIC}_\gamma(\mathbf{E}) - \mathrm{BIC}_\gamma(\mathbf{E}_0) = -2(l_n(\hat{\Theta}(\mathbf{E})) - l_n(\hat{\Theta}(\mathbf{E}_0))) + 4(1+\gamma_0)(|\mathbf{E}| - |\mathbf{E}_0|)\log p > 0.$$

If instead $\mathbf{E} \in \mathcal{E}_1$, then (by Theorem 1, since $|\mathbf{E}_0| \leq q$):

$$\mathrm{BIC}_\gamma(\mathbf{E}) - \mathrm{BIC}_\gamma(\mathbf{E}_0) = -2(l_n(\hat{\Theta}(\mathbf{E})) - l_n(\hat{\Theta}(\mathbf{E}_0))) + 4(1+\gamma_0)(|\mathbf{E}| - |\mathbf{E}_0|)\log p > 0.$$

Therefore, for any $\mathbf{E} \in \mathcal{E}\backslash\{\mathbf{E}_0\}$, $\mathrm{BIC}_\gamma(\mathbf{E}) > \mathrm{BIC}_\gamma(\mathbf{E}_0)$, which yields the desired result. □

Some details on the proofs of Theorems 1 and 2 are given in the Appendix in Section 5.

3 Simulations

In this section, we demonstrate that the EBIC with positive γ indeed leads to better model selection properties in practically relevant settings. We let n grow, set $p \propto n^\kappa$ for various values of κ, and apply the EBIC with $\gamma \in \{0, 0.5, 1\}$ similarly to the choice made in the regression context by [3]. As mentioned in the introduction, we first use the graphical lasso of [2] (as implemented in the 'glasso' package for R) to define a small set of models to consider (details given below). From the selected set we choose the model with the lowest EBIC. This is repeated for 100 trials for each combination of values of n, p, γ in each scaling scenario. For each case, the average positive selection rate (PSR) and false discovery rate (FDR) are computed.

We recall that the graphical lasso places an ℓ_1 penalty on the inverse covariance matrix. Given a penalty $\rho \geq 0$, we obtain the estimate

$$\hat{\Theta}_\rho = \arg\min_{\Theta} -l_n(\Theta) + \rho\|\Theta\|_1. \tag{4}$$

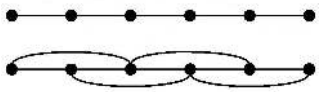

Figure 1: The chain (top) and the 'double chain' (bottom) on 6 nodes.

(Here we may define $\|\Theta\|_1$ as the sum of absolute values of all entries, or only of off-diagonal entries; both variants are common). The ℓ_1 penalty promotes zeros in the estimated inverse covariance matrix $\hat{\Theta}_\rho$; increasing the penalty yields an increase in sparsity. The 'glasso path', that is, the set of models recovered over the full range of penalties $\rho \in [0, \infty)$, gives a small set of models which, roughly, include the 'best' models at various levels of sparsity. We may therefore apply the EBIC to this manageably small set of models (without further restriction to decomposable models). Consistency results on the graphical lasso require the penalty ρ to satisfy bounds that involve measures of regularity in the unknown matrix Θ_0; see [7]. Minimizing the EBIC can be viewed as a data-driven method of tuning ρ, one that does not require creation of test data.

While cross-validation does not generally have consistency properties for model selection (see [9]), it is nevertheless interesting to compare our method to cross-validation. For the considered simulated data, we start with the set of models from the 'glasso path', as before, and then perform 100-fold cross-validation. For each model and each choice of training set and test set, we fit the model to the training set and then evaluate its performance on each sample in the test set, by measuring error in predicting each individual node conditional on the other nodes and then taking the sum of the squared errors. We note that this method is computationally much more intensive than the BIC or EBIC, because models need to be fitted many more times.

3.1 Design

In our simulations, we examine the EBIC as applied to the case where the graph is a chain with node j being connected to nodes $j-1, j+1$, and to the 'double chain', where node j is connected to nodes $j-2, j-1, j+1, j+2$. Figure 1 shows examples of the two types of graphs, which have on the order of p and $2p$ edges, respectively. For both the chain and the double chain, we investigate four different scaling scenarios, with the exponent κ selected from $\{0.5, 0.9, 1, 1.1\}$. In each scenario, we test $n = 100, 200, 400, 800$, and define $p \propto n^\kappa$ with the constant of proportionality chosen such that $p = 10$ when $n = 100$ for better comparability.

In the case of a chain, the true inverse covariance matrix Θ_0 is tridiagonal with all diagonal entries $(\Theta_0)_{j,j}$ set equal to 1, and the entries $(\Theta_0)_{j,j+1} = (\Theta_0)_{j+1,j}$ that are next to the main diagonal equal to 0.3. For the double chain, Θ_0 has all diagonal entries equal to 1, the entries next to the main diagonal are $(\Theta_0)_{j,j+1} = (\Theta_0)_{j+1,j} = 0.2$ and the remaining non-zero entries are $(\Theta_0)_{j,j+2} = (\Theta_0)_{j+2,j} = 0.1$. In both cases, the choices result in values for θ_0, $\sigma^2_{\max}$ and $\lambda_{\max}$ that are bounded uniformly in the matrix size p.

For each data set generated from $N(0, \Theta_0^{-1})$, we use the 'glasso' package [2] in R to compute the 'glasso path'. We choose 100 penalty values ρ which are logarithmically evenly spaced between $\rho_{\max}$ (the smallest value which will result in a no-edge model) and $\rho_{\max}/100$. At each penalty value ρ, we compute $\hat{\Theta}_\rho$ from (4) and define the model $\mathbf{E}_\rho$ based on this estimate's support. The R routine also allows us to compute the unpenalized maximum likelihood estimate $\hat{\Theta}(\mathbf{E}_\rho)$. We may then readily compute the EBIC from (1). There is no guarantee that this procedure will find the model with the lowest EBIC along the full 'glasso path', let alone among the space of all possible models of size $\leq q$. Nonetheless, it serves as a fast way to select a model without any manual tuning.

3.2 Results

Chain graph: The results for the chain graph are displayed in Figure 2. The figure shows the positive selection rate (PSR) and false discovery rate (FDR) in the four scaling scenarios. We observe that, for the larger sample sizes, the recovery of the non-zero coefficients is perfect or nearly perfect for all three values of γ; however, the FDR rate is noticeably better for the positive values of γ, especially

for higher scaling exponents κ. Therefore, for moderately large n, the EBIC with $\gamma = 0.5$ or $\gamma = 1$ performs very well, while the ordinary BIC_0 produces a non-trivial amount of false positives. For 100-fold cross-validation, while the PSR is initially slightly higher, the growing FDR demonstrates the extreme inconsistency of this method in the given setting.

Double chain graph: The results for the double chain graph are displayed in Figure 3. In each of the four scaling scenarios for this case, we see a noticeable decline in the PSR as γ increases. Nonetheless, for each value of γ, the PSR increases as n and p grow. Furthermore, the FDR for the ordinary BIC_0 is again noticeably higher than for the positive values of γ, and in the scaling scenarios $\kappa \geq 0.9$, the FDR for BIC_0 is actually increasing as n and p grow, suggesting that asymptotic consistency may not hold in these cases, as is supported by our theoretical results. 100-fold cross-validation shows significantly better PSR than the BIC and EBIC methods, but the FDR is again extremely high and increases quickly as the model grows, which shows the unreliability of cross-validation in this setting. Similarly to what Chen and Chen [3] conclude for the regression case, it appears that the EBIC with parameter $\gamma = 0.5$ performs well. Although the PSR is necessarily lower than with $\gamma = 0$, the FDR is quite low and decreasing as n and p grow, as desired.

For both types of simulations, the results demonstrate the trade-off inherent in choosing γ in the finite (non-asymptotic) setting. For low values of γ, we are more likely to obtain a good (high) positive selection rate. For higher values of γ, we are more likely to obtain a good (low) false discovery rate. (In the Appendix, this corresponds to assumptions (5) and (6)). However, asymptotically, the conditions (3) guarantee consistency, meaning that the trade-off becomes irrelevant for large n and p. In the finite case, $\gamma = 0.5$ seems to be a good compromise in simulations, but the question of determining the best value of γ in general settings is an open question. Nonetheless, this method offers guaranteed asymptotic consistency for (known) values of γ depending only on n and p.

4 Discussion

We have proposed the use of an extended Bayesian information criterion for multivariate data generated by sparse graphical models. Our main result gives a specific scaling for the number of variables p, the sample size n, the bound on the number of edges q, and other technical quantities relating to the true model, which will ensure asymptotic consistency. Our simulation study demonstrates the the practical potential of the extended BIC, particularly as a way to tune the graphical lasso. The results show that the extended BIC with positive γ gives strong improvement in false discovery rate over the classical BIC, and even more so over cross-validation, while showing comparable positive selection rate for the chain, where all the signals are fairly strong, and noticeably lower, but steadily increasing, positive selection rate for the double chain with a large number of weaker signals.

5 Appendix

We now sketch proofs of non-asymptotic versions of Theorems 1 and 2, which are formulated as Theorems 3 and 4. We also give a non-asymptotic formulation of the Main Theorem; see Theorem 5. In the non-asymptotic approach, we treat all quantities as fixed (e.g. n, p, q, etc.) and state precise assumptions on those quantities, and then give an explicit lower bound on the probability of the extended BIC recovering the model $\mathbf{E}_0$ exactly. We do this to give an intuition for the magnitude of the sample size n necessary for a good chance of exact recovery in a given setting but due to the proof techniques, the resulting implications about sample size are extremely conservative.

5.1 Preliminaries

We begin by stating two lemmas that are used in the proof of the main result, but are also more generally interesting as tools for precise bounds on Gaussian and chi-square distributions. First, Cai [10, Lemma 4] proves the following chi-square bound. For any $n \geq 1, \lambda > 0$,

$$P\{\chi_n^2 > n(1+\lambda)\} \leq \frac{1}{\lambda\sqrt{\pi n}} e^{-\frac{n}{2}(\lambda - \log(1+\lambda))}.$$

We can give an analagous left-tail upper bound. The proof is similar to Cai's proof and omitted here. We will refer to these two bounds together as (CSB).

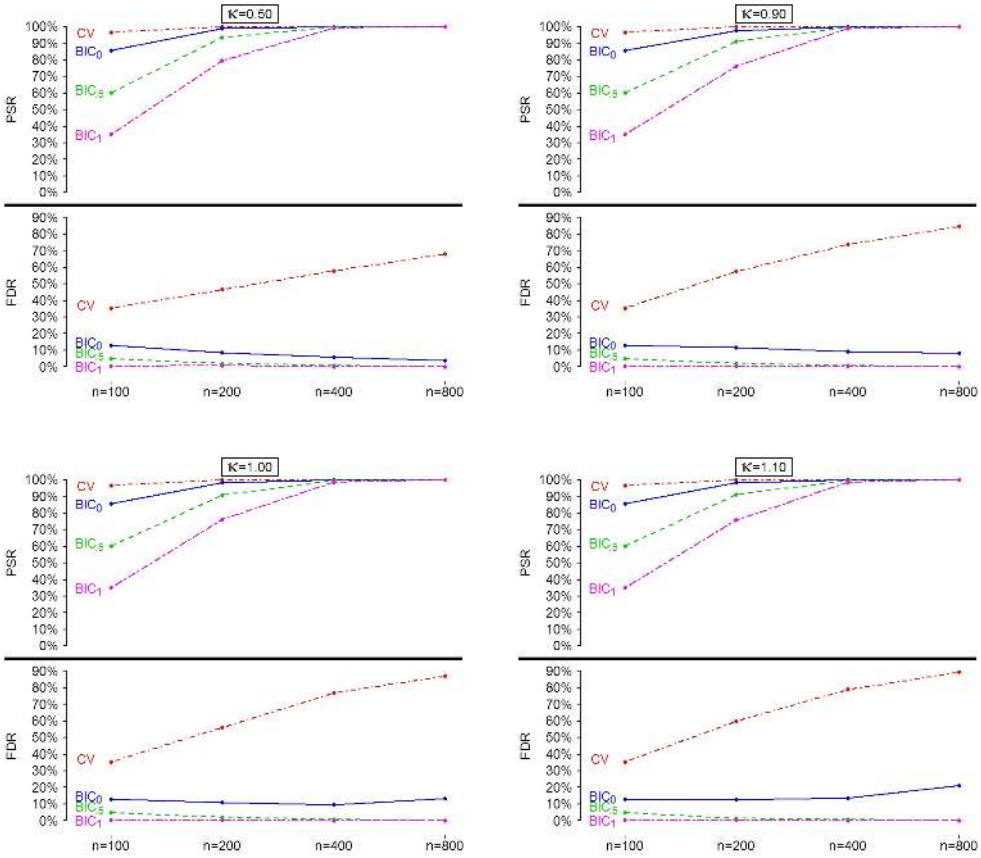

Figure 2: Simulation results when the true graph is a chain.

Lemma 1. *For any $\lambda > 0$, for n such that $n \geq 4\lambda^{-2} + 1$,*

$$P\{\chi_n^2 < n(1-\lambda)\} \leq \frac{1}{\lambda\sqrt{\pi(n-1)}} e^{\frac{n-1}{2}(\lambda + \log(1-\lambda))}.$$

Second, we give a distributional result about the sample correlation when sampling from a bivariate normal distribution.

Lemma 2. *Suppose $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent draws from a bivariate normal distribution with zero mean, variances equal to one and covariance ρ. Then the following distributional equivalence holds, where A and B are independent χ_n^2 variables:*

$$\sum_{i=1}^{n}(X_i Y_i - \rho) \overset{\mathcal{D}}{=} \frac{1+\rho}{2}(A-n) - \frac{1-\rho}{2}(B-n).$$

Proof. Let $A_1, B_1, A_2, B_2, \ldots, A_n, B_n$ be independent standard normal random variables. Define:

$$X_i = \sqrt{\frac{1+\rho}{2}} A_i + \sqrt{\frac{1-\rho}{2}} B_i;\ Y_i = \sqrt{\frac{1+\rho}{2}} A_i - \sqrt{\frac{1-\rho}{2}} B_i;\ A = \sum_{i=1}^{n} A_i^2;\ B = \sum_{i=1}^{n} B_i^2.$$

Then the variables $X_1, Y_1, X_2, Y_2, \ldots, X_n, Y_n$ have the desired joint distribution, and A, B are independent χ_n^2 variables. The claim follows from writing $\sum_i X_i Y_i$ in terms of A and B. □

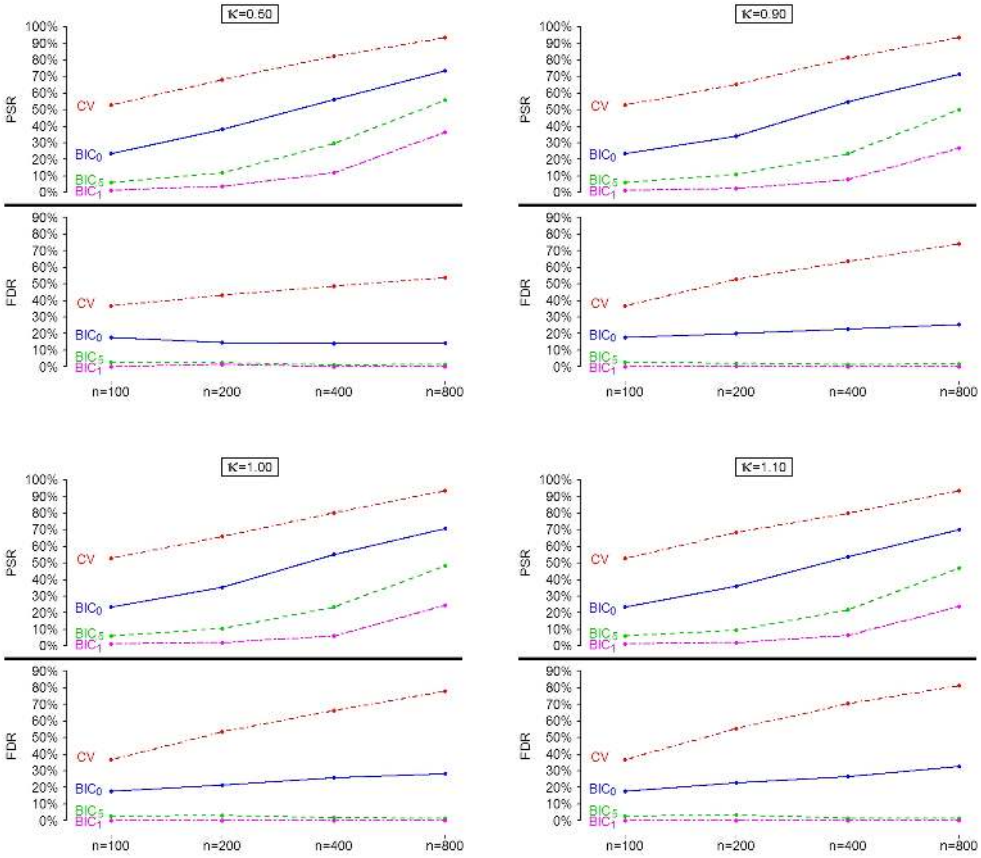

Figure 3: Simulation results when the true graph is a 'double chain'.

5.2 Non-asymptotic versions of the theorems

We assume the following two conditions, where $\epsilon_0, \epsilon_1 > 0$, $C \geq \sigma_{\max}^2 \lambda_{\max}$, $\kappa = \log_n p$, and $\gamma_0 = \gamma - (1 - \frac{1}{4\kappa})$:

$$\frac{(p+2q)\log p}{n} \times \frac{\lambda_{\max}^2}{\theta_0^2} \leq \frac{1}{3200 \max\{1+\gamma_0, \left(1+\frac{\epsilon_1}{2}\right) C^2\}} \tag{5}$$

$$2(\sqrt{1+\gamma_0} - 1) - \frac{\log\log p + \log(4\sqrt{1+\gamma_0}) + 1}{2\log p} \geq \epsilon_0 \tag{6}$$

Theorem 3. *Suppose assumption (5) holds. Then with probability at least* $1 - \frac{1}{\sqrt{\pi \log p}} p^{-\epsilon_1}$*, for all* $\mathbf{E} \not\supset \mathbf{E}_0$ *with* $|\mathbf{E}| \leq q$*,*

$$l_n(\Theta_0) - l_n(\hat{\Theta}(\mathbf{E})) > 2q(\log p)(1+\gamma_0).$$

Proof. We sketch a proof along the lines of the proof of Theorem 2 in [6], using Taylor series centered at the true Θ_0 to approximate the likelihood at $\hat{\Theta}(\mathbf{E})$. The score and the negative Hessian of the log-likelihood function in (2) are

$$s_n(\Theta) = \frac{d}{d\Theta} l_n(\Theta) = \frac{n}{2}\left(\Theta^{-1} - S\right), \qquad H_n(\Theta) = -\frac{d}{d\Theta} s_n(\Theta) = \frac{n}{2}\Theta^{-1} \otimes \Theta^{-1}.$$

Here, the symbol $\otimes$ denotes the Kronecker product of matrices. Note that, while we require Θ to be symmetric positive definite, this is not reflected in the derivatives above. We adopt this convention for the notational convenience in the sequel.

Next, observe that $\hat{\Theta}(\mathbf{E})$ has support on $\Delta \cup \mathbf{E}_0 \cup \mathbf{E}$, and that by definition of θ_0, we have the lower bound $|\hat{\Theta}(\mathbf{E}) - \Theta_0|_F \geq \theta_0$ in terms of the Frobenius norm. By concavity of the log-likelihood function, it suffices to show that the desired inequality holds for all Θ with support on $\Delta \cup \mathbf{E}_0 \cup \mathbf{E}$ with $|\Theta - \Theta_0|_F = \theta_0$. By Taylor expansion, for some $\tilde{\Theta}$ on the path from Θ_0 to Θ, we have:

$$l_n(\Theta) - l_n(\Theta_0) = \mathrm{vec}(\Theta - \Theta_0)^T s_n(\Theta_0) - \frac{1}{2}\mathrm{vec}(\Theta - \Theta_0)^T H_n(\tilde{\Theta})\mathrm{vec}(\Theta - \Theta_0).$$

Next, by (CSB) and Lemma 2, with probability at least $1 - \frac{1}{\sqrt{\pi \log p}} e^{-\epsilon_1 \log p}$, the following bound holds for all edges $\mathbf{e}$ in the complete graph (we omit the details):

$$(s_n(\Theta_0))_{\mathbf{e}}^2 \leq 6\sigma_{\max}^4(2 + \epsilon_1)n \log p.$$

Now assume that this bound holds for all edges. Fix some $\mathbf{E}$ as above, and fix Θ with support on $\Delta \cup \mathbf{E}_0 \cup \mathbf{E}$, with $|\Theta - \Theta_0| = \theta_0$. Note that the support has at most $(p + 2q)$ entries. Therefore,

$$|\mathrm{vec}(\Theta - \Theta_0)^T s_n(\Theta_0)|^2 \leq \theta_0^2(p + 2q) \times 6\sigma_{\max}^4(2 + \epsilon_1)n \log p.$$

Furthermore, the eigenvalues of Θ are bounded by $\lambda_{\max} + \theta_0 \leq 2\lambda_{\max}$, and so by properties of Kronecker products, the minimum eigenvalue of $H_n(\tilde{\Theta})$ is at least $\frac{n}{2}(2\lambda_{\max})^{-2}$. We conclude that

$$l_n(\Theta) - l_n(\Theta_0) \leq \sqrt{\theta_0^2(p + 2q) \times 6\sigma_{\max}^4(2 + \epsilon_1)n \log p} - \frac{1}{2}\theta_0^2 \times \frac{n}{2}(2\lambda_{\max})^{-2}.$$

Combining this bound with our assumptions above, we obtain the desired result. □

Theorem 4. *Suppose additionally that assumption (6) holds (in particular, this implies that $\gamma > 1 - \frac{1}{4\kappa}$). Then with probability at least $1 - \frac{1}{4\sqrt{\pi}\log p}\frac{p^{-\epsilon_0}}{1-p^{-\epsilon_0}}$, for all decomposable models $\mathbf{E}$ such that $\mathbf{E} \supsetneq \mathbf{E}_0$ and $|\mathbf{E}| \leq q$,*

$$l_n(\hat{\Theta}(\mathbf{E})) - l_n(\hat{\Theta}(\mathbf{E}_0)) < 2(1 + \gamma_0)(|\mathbf{E}| - |\mathbf{E}_0|)\log p.$$

Proof. First, fix a single such model $\mathbf{E}$, and define $m = |\mathbf{E}| - |\mathbf{E}_0|$. By [8, 11], $l_n(\hat{\Theta}(\mathbf{E})) - l_n(\hat{\Theta}(\mathbf{E}_0))$ is distributed as $-\frac{n}{2}\log\left(\prod_{i=1}^m B_i\right)$, where $B_i \sim Beta(\frac{n-c_i}{2}, \frac{1}{2})$ are independent random variables and the constants $c_1, \ldots, c_m$ are bounded by 1 less than the maximal clique size of the graph given by model $\mathbf{E}$, implying $c_i \leq \sqrt{2q}$ for each i. Also shown in [8] is the stochastic inequality $-\log(B_i) \leq \frac{1}{n-c_i-1}\chi_1^2$. It follows that, stochastically,

$$l_n(\hat{\Theta}(\mathbf{E})) - l_n(\hat{\Theta}(\mathbf{E}_0)) \leq \frac{n}{2} \times \frac{1}{n - \sqrt{2q} - 1}\chi_m^2.$$

Finally, combining the assumptions on n, p, q and the (CSB) inequalities, we obtain:

$$P\{l_n(\hat{\Theta}(\mathbf{E})) - l_n(\hat{\Theta}(\mathbf{E}_0)) \geq 2(1 + \gamma_0)m\log(p)\} \leq \frac{1}{4\sqrt{\pi}\log p} e^{-\frac{m}{2}(4(1+\frac{\epsilon_0}{2})\log p)}.$$

Next, note that the number of models $|\mathbf{E}|$ with $\mathbf{E} \supset \mathbf{E}_0$ and $|\mathbf{E}| - |\mathbf{E}_0| = m$ is bounded by p^{2m}. Taking the union bound over all choices of m and all choices of $\mathbf{E}$ with that given m, we obtain that the desired result holds with the desired probability. □

We are now ready to give a non-asymptotic version of the Main Theorem. For its proof apply the union bound to the statements in Theorems 3 and 4, as in the asymptotic proof given in section 2.

Theorem 5. *Suppose assumptions (5) and (6) hold. Let $\mathcal{E}$ be the set of subsets $\mathbf{E}$ of edges between the p nodes, satisfying $|\mathbf{E}| \leq q$ and representing a decomposable model. Then it holds with probability at least $1 - \frac{1}{4\sqrt{\pi}\log p}\frac{p^{-\epsilon_0}}{1-p^{-\epsilon_0}} - \frac{1}{\sqrt{\pi \log p}}p^{-\epsilon_1}$ that*

$$\mathbf{E}_0 = \arg\min_{\mathbf{E} \in \mathcal{E}} \mathrm{BIC}_\gamma(\mathbf{E}).$$

That is, the extended BIC with parameter γ selects the smallest true model.

Finally, we note that translating the above to the asymptotic version of the result is simple. If the conditions (3) hold, then for sufficiently large n (and thus sufficiently large p), assumptions (5) and (6) hold. Furthermore, although we may not have the exact equality $\kappa = \log_n p$, we will have $\log_n p \to \kappa$; this limit will be sufficient for the necessary inequalities to hold for sufficiently large n. The proofs then follow from the non-asymptotic results.

References

[1] Steffen L. Lauritzen. *Graphical models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press Oxford University Press, New York, 1996. Oxford Science Publications.

[2] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[3] Jiahua Chen and Zehua Chen. Extended Bayesian information criterion for model selection with large model space. *Biometrika*, 95:759–771, 2008.

[4] Gideon Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 1978.

[5] Malgorzata Bogdan, Jayanta K. Ghosh, and R. W. Doerge. Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci. *Genetics*, 167:989–999, 2004.

[6] Jiahua Chen and Zehua Chen. Extended BIC for small-n-large-p sparse GLM. Preprint.

[7] Pradeep Ravikumar, Martin J. Wainwright, Garvesh Raskutti, and Bin Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. arXiv:0811.3628, 2008.

[8] B. T. Porteous. Stochastic inequalities relating a class of log-likelihood ratio statistics to their asymptotic χ^2 distribution. *Ann. Statist.*, 17(4):1723–1734, 1989.

[9] Jun Shao. Linear model selection by cross-validation. *J. Amer. Statist. Assoc.*, 88(422):486–494, 1993.

[10] T. Tony Cai. On block thresholding in wavelet regression: adaptivity, block size, and threshold level. *Statist. Sinica*, 12(4):1241–1273, 2002.

[11] P. Svante Eriksen. Tests in covariance selection models. *Scand. J. Statist.*, 23(3):275–284, 1996.