University of Groningen

Selecting the tuning parameter in penalized Gaussian graphical models

Abbruzzo, Antonino; Vujacic, Ivan; Mineo, Angelo M.; Wit, Ernst C.

Published in:
Statistics and Computing

DOI:
[10.1007/s11222-018-9823-5](https://doi.org/10.1007/s11222-018-9823-5)

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

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2019

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

Citation for published version (APA):

Abbruzzo, A., Vujacic, I., Mineo, A. M., & Wit, E. C. (2019). Selecting the tuning parameter in penalized Gaussian graphical models. *Statistics and Computing*, 29(3), 559-569. <https://doi.org/10.1007/s11222-018-9823-5>

Copyright

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

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

Take-down policy

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

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

Selecting the tuning parameter in penalized Gaussian graphical models

Antonino Abbruzzo¹ · Ivan Vujačić² · Angelo M. Mineo¹ · Ernst C. Wit^{3,4}

Received: 26 September 2017 / Accepted: 3 July 2018 / Published online: 9 July 2018
© Springer Science+Business Media, LLC, part of Springer Nature 2018

Abstract

Penalized inference of Gaussian graphical models is a way to assess the conditional independence structure in multivariate problems. In this setting, the conditional independence structure, corresponding to a graph, is related to the choice of the tuning parameter, which determines the model complexity or degrees of freedom. There has been little research on the degrees of freedom for penalized Gaussian graphical models. In this paper, we propose an estimator of the degrees of freedom in ℓ_1 -penalized Gaussian graphical models. Specifically, we derive an estimator inspired by the generalized information criterion and propose to use this estimator as the bias term for two information criteria. We called these tuning parameter selectors GAIC and GBIC. These selectors can be used to choose the tuning parameter, i.e., the optimal tuning parameter is the minimizer of GAIC or GBIC. A simulation study shows that GAIC tends to improve the performance of both AIC-type and CV-type model selectors, in terms of estimation quality (entropy loss function) in high-dimensional setting. Moreover, GBIC model selector improves the performance of both BIC-type and CV-type model selectors, in terms of support recovery (F-score). A data analysis shows that GBIC selects a tuning parameter that produces a sparser graph with respect to BIC and a CV-type model selector (KLCV).

Keywords Penalized likelihood · Kullback–Leibler divergence · Model complexity · Model selection · Generalized information criterion

1 Introduction

A graphical model is composed of nodes that represent random variables and edges that represent Markov dependencies between variables. The Gaussian graphical model (GGM) is a graphical model in which the nodes are jointly normally

distributed. In GGMs, conditional dependencies are identified in the inverse covariance matrix, called the precision matrix. Nonzero elements in the precision matrix correspond to conditionally dependent variables. One main goal in GGM inference is to estimate the precision matrix [see Edwards (2000), Lauritzen (1996), Rue and Held (2005), and Whittaker (2009)].

Let $\mathbf{Y}$ be a random variable in $\mathbb{R}^p$, and suppose we have n multivariate observations of dimension p from $\mathcal{N}(\mathbf{0}, \mathbf{\Omega}^{-1})$, where $\mathbf{\Omega} = (\omega_{ij})_{i,j}$ is the precision matrix. Let $\mathbf{Y}$ be a $n \times p$ matrix of observations. The log-likelihood of observation $\mathbf{y}_k$ is

$$\ell_k(\mathbf{\Omega}) = \{\log |\mathbf{\Omega}| - \text{tr}(\mathbf{\Omega} \mathbf{S}_k)\}/2, \quad (1)$$

for $k = 1, \dots, n$, where $\mathbf{S}_k = \mathbf{y}_k \mathbf{y}_k^\top$. Let $\mathbf{S} = \sum_{k=1}^n \mathbf{S}_k/n$ be the empirical covariance matrix. The log-likelihood for n independent and identically distributed samples is $\ell(\mathbf{\Omega}) = \sum_{i=1}^n \ell_k(\mathbf{\Omega})$, and the penalized maximum log-likelihood estimator (PMLE), with penalty function $p_\lambda(\cdot)$, is obtained by maximizing

This research was supported in part by COSTNET (CA15109).

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s11222-018-9823-5>) contains supplementary material, which is available to authorized users.

✉ Antonino Abbruzzo
antonino.abbruzzo@unipa.it

¹ Department of Economics, Business and Statistics, University of Palermo, Viale delle Scienze Ed. 13, 90128 Palermo, Italy

² PRA Health Sciences, Early Development Services, Paterswoldseweg 811, 9728 BM Groningen, The Netherlands

³ Johann Bernoulli Institute, University of Groningen, Nijenborgh 9, 9747 AG Groningen, The Netherlands

⁴ Institute of Computation Science, USI, Via Buffi 13, 6qoo Lugano, Switzerland

$$\ell_{\lambda}(\mathbf{\Omega}) = \log |\mathbf{\Omega}| - \text{tr}(\mathbf{\Omega}\mathbf{S}) - \sum_{i=1}^p \sum_{j=1}^p p_{\lambda}(|\omega_{ij}|), \quad (2)$$

over the set of positive definite matrices, where λ is the tuning parameter. The resulting PMLE is denoted as $\hat{\mathbf{\Omega}}_{\lambda}$.

Several penalty functions have been proposed in the literature. The L_1 -norm (lasso penalty) is probably the most known Cai et al. (2011), Friedman et al. (2008). Other penalties include the weighted L_1 -norm (adaptive lasso) and the Smoothly Clipped Absolute Deviation (SCAD) Fan et al. (2009). The L_1 -norm is convex and leads to a desirable convex optimization problem if the log-likelihood function is convex, which is the case for GGMs. Whereas it has been shown that the lasso penalty produces biases even in the simple regression setting Fan and Li (2001), the SCAD penalty and the adaptive lasso reduce this bias.

Optimizing (2) with a L_1 -norm penalty can be solved by several algorithms. For example, Yuan and Lin (2007) solved this optimization problem using the MAXDET algorithm. d'Aspremont et al. (2008) proposed two efficient first-order numerical algorithms with low memory requirements using semidefinite programming algorithms, which obey the positive definiteness constraint of the precision matrix. Cai et al. (2011) proposed a constraint L_1 optimization method, called CLIME. Friedman et al. (2008) proposed the graphical lasso algorithm, which uses a coordinate descent procedure.

The theoretical properties of the estimator $\hat{\mathbf{\Omega}}_{\lambda}$ are well studied [e.g., Friedman et al. (2008), Lam and Fan (2009), Rothman et al. (2008), Yuan and Lin (2007)]. In particular, Rothman et al. (2008) and Fan et al. (2009) showed that the Frobenius norm between the inverse covariance matrix and its L_1 penalized likelihood estimator is $O_p(\sqrt{(E \log p/n)})$, where E is the number of the nonzero elements of the inverse of the correlation matrix. Consequently, a sparse inverse covariance matrix is highly estimable.

The importance of correctly identifying the precision matrix $\mathbf{\Omega}$ is due to its large use. For example, several multivariate analyses, such as principal component analysis or linear discriminant analysis, require the estimation of the covariance matrix or its inverse. In microarray analysis, the precision matrix may help in understanding the relationships between the expression levels of a large set of genes.

In order to identify a precision matrix $\hat{\mathbf{\Omega}}_{\lambda}$ (as close as possible to $\mathbf{\Omega}$), we have to select a tuning parameter λ . From an asymptotic point of view, it has been shown that graphical lasso and the CLIME estimator Cai et al. (2011) can achieve model selection consistency for sparse precision matrix estimation, when the tuning parameter is chosen properly. In other words, both graphical lasso and CLIME estimators can fully recover the support of the unknown precision matrix. In the finite sample scenario, the performance of the penalized likelihood estimator $\hat{\mathbf{\Omega}}_{\lambda}$, in minimizing the Kullback–Leibler

divergence, can be related to the bias term $E[\ell(\mathbf{\Omega}) - \ell(\hat{\mathbf{\Omega}}_{\lambda})]$. Yuan and Lin (2007) suggests to consider as bias term the number of nonzero elements in the precision matrix. This estimator was used in Gao et al. (2012), in combination with a Bayesian information criterion, and in Foygel and Drton (2010), with an extended Bayesian information criterion, to select the tuning parameter. Classical criteria such as AIC and BIC tend to select a model with many spurious covariates when the number of variables is much greater than the number of observations. This phenomenon has been observed by Broman and Speed (2002), Siegmund (2004), Bogdan et al. (2004), and Chen and Chen (2008). Cross-validation (CV) may be a valid alternative in this regard, but it is computationally expensive. So, various approximations have been proposed Lian (2011), Vujačić et al. (2015), which makes its calculation computationally feasible even in high-dimensional settings.

We propose an estimator for the bias term (so-called degrees of freedom) which is inspired by the generalized information criterion (GIC) proposed in Konishi and Kitagawa (1996). Nishii (1984) had proposed a different GIC than Konishi and Kitagawa (1996). See also Zhang et al. (2010). The GIC formula from Konishi and Kitagawa (1996) is not directly applicable to problems defined by (2) that give sparse solutions. This is because penalty functions satisfying the conditions of sparsity and continuity must be singular at the origin Fan and Li (2001). Therefore, we propose an approximate solution in order to overcome the problem of differentiability. The approximate bias term is used in AIC-type or BIC-type selection methods of the regularization parameter, and the resulting model selectors are called GAIC and GBIC, respectively. These selectors can be used to choose the tuning parameter, i.e., the optimal tuning parameter is the minimizer of GAIC or GBIC.

A simulation study shows that GAIC tends to improve the performance of both AIC-type and CV-type model selectors, in terms of estimation quality (loss entropy) in high-dimensional setting, i.e., when the number of variables is much greater than the number of observations. GBIC model selector improves the performance of both BIC-type and CV-type model selectors, in terms of support recovery (F-score). Both the estimator seems to be consistent in terms of Frobenius norm. Only for GBIC, F-score approaches to 1 with increasing sample size. Moreover, we considered three types of penalized Gaussian graphical models: the graphical lasso, the adaptive lasso, and the SCAD. The simulation study shows that the adaptive lasso improves the quality of the estimator both in terms of Frobenius norm and F-score. Finally, a data analysis shows that GBIC selects a tuning parameter that produces a sparser graph with respect to BIC and KLCV.

The rest of the paper is organized as follows. In Sect. 2, we show the proposed model selectors and derive the generalized information criterion for Gaussian graphical models.

In Sect. 3, we propose an estimator of the bias term which is derived from an approximation of the generalized information criterion for the ℓ_1 -penalized maximum likelihood estimator. Numerical results are presented in Sect. 4. In Sect. 5, the GBIC is used in reconstruction of a biological network. We conclude with some remarks in Sect. 6. “Appendix A” contains proofs and auxiliary materials.

2 Generalized information criterion

In this section, we define the generalized information criteria (GIC) for model selection in p -dimensional Gaussian graphical models. The bias correction term, denoted as $\text{df}(\cdot)$, derived from the GIC is used in AIC-type and BIC-type model selection methods, to define two tuning parameter selectors:

$$\text{GAIC}(\lambda) = n\{\log |\hat{\boldsymbol{\Omega}}_\lambda| - \text{tr}(\hat{\boldsymbol{\Omega}}_\lambda \mathbf{S})\} + 2\text{df}(\lambda), \quad (3)$$

and

$$\text{GBIC}(\lambda) = n\{\log |\hat{\boldsymbol{\Omega}}_\lambda| - \text{tr}(\hat{\boldsymbol{\Omega}}_\lambda \mathbf{S})\} + \log(n)\text{df}(\lambda). \quad (4)$$

These selector estimators, derived in Sect. 3, can be used to choose the optimal λ , i.e., the minimizer either of (3) or (4).

2.1 Generalized information criteria for Gaussian graphical models

The generalized information criterion (GIC) Konishi and Kitagawa (1996) provides an estimate of Kullback–Leibler divergence for a wide class of estimators which includes M -estimators. The maximum likelihood estimator for the log-likelihood in (1) is an M -estimator which is defined as a solution of the system of equations

$$\sum_{k=1}^n \boldsymbol{\psi}(\mathbf{y}_k, \boldsymbol{\Omega}) = \mathbf{0}, \quad (5)$$

where $\boldsymbol{\psi}(\mathbf{y}_k, \boldsymbol{\Omega}) = \text{vec} D\ell_k(\boldsymbol{\Omega})$, ℓ_k is defined in (1), D is the partial derivative operator, $D\ell_k(\boldsymbol{\Omega})$ is the matrix of partial derivatives of ℓ_k with respect to ω_{ij} , and vec is the operator that transforms a matrix into a vector by stacking its columns one underneath the other. The GIC for M -estimator $\hat{\boldsymbol{\Omega}}$, where $\hat{\boldsymbol{\Omega}}$ is the solution of (5), is defined by (Konishi and Kitagawa 2008, p. 117):

$$\text{GIC} = -2 \sum_{k=1}^n \ell_k(\hat{\boldsymbol{\Omega}}) + 2\text{tr}(\mathbf{R}^{-1}\mathbf{Q}), \quad (6)$$

where $\mathbf{R}$ and $\mathbf{Q}$ are square matrices of order p^2 given by

$$\begin{aligned} \mathbf{R} &= -\frac{1}{n} \sum_{k=1}^n \{D\boldsymbol{\psi}(\mathbf{y}_k, \hat{\boldsymbol{\Omega}})\}^\top, \\ \mathbf{Q} &= \frac{1}{n} \sum_{k=1}^n \boldsymbol{\psi}(\mathbf{y}_k, \hat{\boldsymbol{\Omega}}) D\ell_k(\hat{\boldsymbol{\Omega}}). \end{aligned} \quad (7)$$

By using matrix differential calculus and some algebra, we obtain:

$$\begin{aligned} \mathbf{R}^{-1} &= 2\hat{\boldsymbol{\Omega}} \otimes \hat{\boldsymbol{\Omega}}, \\ \mathbf{Q} &= \frac{1}{4n} \sum_{k=1}^n \text{vec} \mathbf{S}_k \text{vec} \mathbf{S}_k^\top - \frac{1}{4} \text{vec} \mathbf{S} \text{vec} \mathbf{S}^\top, \end{aligned} \quad (8)$$

where, given two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{p \times p}$, the Kronecker product $\mathbf{A} \otimes \mathbf{B}$ is defined as the $p^2 \times p^2$ matrix with entries $a_{ij}b_{kl}$ at the position $(i(p-1)+k, j(p-1)+l)$. From (6) and (8), we obtain:

$$\text{tr}(\mathbf{R}^{-1}\mathbf{Q}) = \frac{1}{2n} \sum_{k=1}^n \langle \text{vec} \mathbf{S}_k, \text{vec} \mathbf{S}_k \rangle_{\hat{\mathcal{I}}} - \frac{1}{2} \langle \text{vec} \mathbf{S}, \text{vec} \mathbf{S} \rangle_{\hat{\mathcal{I}}}, \quad (9)$$

where $\hat{\mathcal{I}}^{-1} = \hat{\boldsymbol{\Omega}} \otimes \hat{\boldsymbol{\Omega}}$, and $\langle \cdot, \cdot \rangle_{\hat{\mathcal{I}}}$ is the inner product induced by positive definite $p^2 \times p^2$ matrix $\hat{\mathcal{I}}$.

An important fact is that in (8) we have $\mathbf{R}^{-1} = 2\hat{\mathcal{I}}^{-1}$, where $\hat{\mathcal{I}}$ is the Fisher information matrix calculated at $\hat{\boldsymbol{\Omega}}$. Note that in case of Gaussian graphical models the asymptotic covariance matrix of the MLE is not the inverse Fisher information matrix $\mathcal{I}^{-1} = \boldsymbol{\Omega} \otimes \boldsymbol{\Omega}$, but it is equal to

$$\mathcal{AC} = \mathbf{M}_p(\boldsymbol{\Omega} \otimes \boldsymbol{\Omega}), \quad (10)$$

where $\mathbf{M}_p = \frac{1}{2}(\mathbf{I}_{p^2} + \mathbf{K}_p)$, and $\mathbf{I}_{p^2}$ and $\mathbf{K}_p$ are identity matrix and commutation matrix of order p^2 , respectively (see Fried and Vogel (2009)). The commutation matrix $\mathbf{K}_p$ is defined as $p^2 \times p^2$ matrix that has the property $\mathbf{K}_p \text{vec} \mathbf{A} = \text{vec} \mathbf{A}^\top$ for any $p \times p$ matrix $\mathbf{A}$. From (10) by using some algebra, we obtain that

$$\text{cov}(\hat{\omega}_{ij}, \hat{\omega}_{kl}) \rightarrow \frac{1}{2}(\omega_{jl}\omega_{ik} + \omega_{jk}\omega_{il}), \quad n \rightarrow \infty. \quad (11)$$

In view of (11), we would like to write $\text{tr}(\mathbf{R}^{-1}\mathbf{Q})$ in terms of the estimated asymptotic covariance matrix, which is $\widehat{\mathcal{AC}} = \mathbf{M}_p \hat{\mathcal{I}}$ and not of $\hat{\mathcal{I}}$. However, what is more important here is the geometry of the likelihood surface and the Fisher information is a measure of its curvature. Even though $\widehat{\mathcal{AC}} \neq \hat{\mathcal{I}}$, these two matrices yield the same geometry on the vector space of symmetric matrices. Indeed, it can be shown that the following lemma holds.

Lemma 1 Let $\mathbf{A}$, $\mathbf{B}$ be symmetric matrices of order p and $\mathbf{\Omega}$ be a positive definite symmetric matrix of order p . Then, the following identity holds

$$\langle \text{vec} \mathbf{A}, \text{vec} \mathbf{B} \rangle_{\mathbf{\Omega} \otimes \mathbf{\Omega}} = \langle \text{vec} \mathbf{A}, \text{vec} \mathbf{B} \rangle_{\mathbf{M}_p(\mathbf{\Omega} \otimes \mathbf{\Omega})},$$

where $\langle \text{vec} \mathbf{A}, \text{vec} \mathbf{B} \rangle_{\Psi} = \text{vec} \mathbf{B}^T \Psi \text{vec} \mathbf{A}$ is the inner product induced by positive definite $p^2 \times p^2$ matrix Ψ , and $\mathbf{M}_p = \frac{1}{2}(\mathbf{I}_{p^2} + \mathbf{K}_p)$.

See “Appendix A” for the proof. By using Lemma 1, it can be shown that

$$\begin{aligned} \text{tr}(\mathbf{R}^{-1} \mathbf{Q}) &= \frac{1}{2n} \sum_{k=1}^n \langle \text{vec} \mathbf{S}_k, \text{vec} \mathbf{S}_k \rangle_{\widehat{\mathcal{AC}}} \\ &\quad - \frac{1}{2} \langle \text{vec} \mathbf{S}, \text{vec} \mathbf{S} \rangle_{\widehat{\mathcal{AC}}}. \end{aligned} \quad (12)$$

2.2 Relationship between KL and GIC in GGM

The GIC in (6) is related to the Kullback–Leibler (KL) divergence. This divergence is used to measure the “distance” between the model and the true data generating mechanism and as such stands at the basis of several model selection methods. Specifically, assume that the model is given by the distribution $\mathcal{N}(\mathbf{0}, \hat{\mathbf{\Omega}}^{-1})$ and that the true data mechanism is given by the distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Omega}^{-1})$. Then, the KL divergence from the model to the true distribution is

$$\text{KL}(\mathbf{\Omega} | \hat{\mathbf{\Omega}}) = \frac{1}{2} \left\{ \text{tr}(\mathbf{\Omega}^{-1} \hat{\mathbf{\Omega}}) - \log |\mathbf{\Omega}^{-1} \hat{\mathbf{\Omega}}| - p \right\}. \quad (13)$$

This can be written in the scaled form as

$$2n \text{KL}(\mathbf{\Omega} | \hat{\mathbf{\Omega}}) \cong -nl(\hat{\mathbf{\Omega}}) + n \text{tr}\{\hat{\mathbf{\Omega}}(\mathbf{\Omega}^{-1} - \mathbf{S})\}, \quad (14)$$

where $l(\hat{\mathbf{\Omega}}) = \{\log |\mathbf{\Omega}| - \text{tr}(\mathbf{\Omega} \mathbf{S})\}$. From (14), we see that the scaled likelihood $-2l(\hat{\mathbf{\Omega}})$ is a biased estimate of the KL divergence and we can also observe that the (oracle) bias term for a multivariate Gaussian model is given by

$$n \text{tr}\{\hat{\mathbf{\Omega}}(\mathbf{\Omega}^{-1} - \mathbf{S})\} \quad (15)$$

and involves the unknown quantity $\mathbf{\Omega}^{-1}$.

3 Generalized information criterion for penalized Gaussian graphical models

The MLE for GGMs is not uniquely defined for high-dimensional data, and alternative methods mainly based on penalized likelihood have been proposed. In the case of L_1 -norm penalized Gaussian graphical models (PGGMs), the

graphical lasso algorithm (glasso) is probably the most common method for finding conditional independence among large set of random variables. The main property of the L_1 -norm PGGMs is that the maximum penalized likelihood estimator for the precision matrix $\mathbf{\Omega}_\lambda$ can be sparse. The sparsity depends on a tuning parameter λ which we propose to select by using the model selectors in (3) or in (4).

The derivation of $\text{df}(\lambda) = \text{tr}(\mathbf{R}_\lambda^{-1} \mathbf{Q}_\lambda)$ by using GIC formula from Konishi and Kitagawa (1996) for the problems defined by (2) that give sparse solutions is not possible. This is because the formula involves derivatives of the penalty function. As noted in Introduction, penalty functions satisfying the conditions of sparsity and continuity must be singular at the origin.

3.1 Problem of zeros in model selection of PGGMs

In deriving a method for selection of the tuning parameter for L_1 -norm penalized Gaussian graphical models, Lian (2011) proposed to overcome the problem of non-differentiability of the penalty function by ignoring the partial derivatives corresponding to the zero elements in $\hat{\mathbf{\Omega}}_\lambda$. This idea is inspired by Fan and Li (2001) and Wang et al. (2007). However, Lian (2011) derived an erroneous approximation for the inverse of Fisher information matrix that has the full dimension $p^2 \times p^2$. This is not possible, since any matrix of partial derivatives of less than p^2 variables must have a smaller dimension.

To overcome the problem of non-differentiability, we take a different approach. We consider a smooth proxy of the penalized log-likelihood given in (2)

$$\ell_{\lambda, \alpha}(\mathbf{\Omega}) = \log |\mathbf{\Omega}| - \text{tr}(\mathbf{\Omega} \mathbf{S}) - \lambda \sum_{i=1}^p \sum_{j=1}^p |\omega_{ij}|_\alpha,$$

where $|\cdot|_\alpha$ is any convex and differentiable approximation of the L_1 penalty function such that $\lim_{\alpha \rightarrow +\infty} |\omega_{ij}|_\alpha''$ is equal to ∞ for $\omega_{ij} \neq 0$ and to zero for $\omega_{ij} = 0$. One example of a function that satisfies these conditions, and that we will use in this paper, is

$$|\omega_{ij}|_\alpha = \frac{1}{\alpha} [\log\{1 + \exp(-\alpha \omega_{ij})\} + \log\{1 + \exp(\alpha \omega_{ij})\}], \quad (16)$$

see Schmidt et al. (2007). In this way, we obtain what will be referred to as the approximated problem.

For a fixed λ , let $\hat{\mathbf{\Omega}}_{\lambda, \alpha}$ be the solution to the approximated problem. The log-likelihood of the approximated problem is convex, so it follows that

$$\lim_{\alpha \rightarrow +\infty} \hat{\mathbf{\Omega}}_{\lambda, \alpha} = \hat{\mathbf{\Omega}}_\lambda.$$

Let $Hf(\boldsymbol{\Omega})$ be the Hessian matrix for f . A main quantity of interest, which is directly related to $\mathbf{R}$ in (8), for determining choice of tuning parameter λ is the inverse of estimated Hessian matrix given by

$$\begin{aligned}\hat{\mathcal{I}}_{\lambda}(\alpha)^{-1} &:= -\{H\ell_{\lambda,\alpha}(\hat{\boldsymbol{\Omega}}_{\lambda,\alpha})\}^{-1} \\ &= \left\{ \hat{\boldsymbol{\Omega}}_{\lambda,\alpha}^{-1} \otimes \hat{\boldsymbol{\Omega}}_{\lambda,\alpha}^{-1} + \mathbf{D}_{\lambda}(\alpha) \right\}^{-1},\end{aligned}\quad (17)$$

where $\mathbf{D}_{\lambda}(\alpha)$ is the diagonal matrix of order p^2 with the elements on the diagonal equal to $|\hat{\omega}_{\lambda,ij}|''_{\alpha}$, for $i, j = 1, \dots, p$. The following lemma gives the limit of $\hat{\mathcal{I}}_{\lambda}(\alpha)^{-1}$ if the pattern of zeros in $\hat{\boldsymbol{\Omega}}_{\lambda,\alpha}$ is the same as for the MLE and PMLE, respectively.

Lemma 2 *Let $I_{\lambda}(\alpha) \subset \{1, \dots, p^2\}$ be the set of indices of nonzero entries in the diagonal of $\mathbf{D}_{\lambda}(\alpha)$ and I_{λ} be the set of indices of nonzero entries in $\text{vec}\hat{\boldsymbol{\Omega}}_{\lambda}$.*

1. *If there exists α_0 such that $I_{\lambda}(\alpha) = \{1, \dots, p^2\}$, for $\alpha \geq \alpha_0$, then the limit $\lim_{\alpha \rightarrow +\infty} \hat{\mathcal{I}}_{\lambda}(\alpha)^{-1}$, denoted by $\hat{\mathcal{I}}_{\lambda}^{-1}$, exists and*

$$\hat{\mathcal{I}}_{\lambda}^{-1} = \hat{\boldsymbol{\Omega}}_{\lambda} \otimes \hat{\boldsymbol{\Omega}}_{\lambda}.\quad (18)$$

2. *If there exists α_0 such that $I_{\lambda}(\alpha) = I_{\lambda}$ for $\alpha \geq \alpha_0$, then the limit $\lim_{\alpha \rightarrow +\infty} \hat{\mathcal{I}}_{\lambda}(\alpha)^{-1}$, denoted by $\hat{\mathcal{I}}_{\lambda}^{-1}$, exists and*

$$\begin{aligned}\left(\hat{\mathcal{I}}_{\lambda}^{-1}\right)_{i,j} &= 0, \text{ when } i \text{ or } j \notin I_{\lambda}, \\ \hat{\mathcal{I}}_{\lambda}^{-1}[I_{\lambda}, I_{\lambda}] &= \left\{ \left(\hat{\boldsymbol{\Omega}}_{\lambda}^{-1} \otimes \hat{\boldsymbol{\Omega}}_{\lambda}^{-1} \right) [I_{\lambda}, I_{\lambda}] \right\}^{-1}.\end{aligned}\quad (19)$$

See “Appendix A” for the proof.

This result is in accordance with Lam and Fan (2009), where similar matrices appear. The formula obtained by Lian (2011) has the form (18) but should be given by the second term in (19), since the zero elements should be ignored. In fact, expression (11) shows the problem with using formula (18) for the penalized case: If one of $\hat{\omega}_{ij}$, $\hat{\omega}_{kl}$ are equal to zero, then the covariance should be estimated as zero as well, which is not the case. On the other hand, formula (19) that should be used in penalized likelihood approaches is computationally expensive in terms of both memory and floating-point operations.

3.2 Model selection for penalized GGMs

In deriving a criterion that extends the generalized information criterion (GIC) to maximum penalized likelihood estimator in Gaussian graphical models, we assume standard conditions like in Lam and Fan (2009). So, by letting $\lambda \rightarrow 0$ when $n \rightarrow \infty$, we can use formula (9), derived for the maximum likelihood case, as an approximation for

the $\text{df}(\lambda)$ in the penalized case. This means that asymptotically the covariances between zero elements and nonzero elements are equal to zero (c.f. (11)). Thus, to obtain the term $\text{df}(\lambda) = \text{tr}(\mathbf{R}_{\lambda}^{-1} \mathbf{Q}_{\lambda})$ for the PMLE we substitute $\hat{\boldsymbol{\Omega}}_{\lambda}$ in (12) and also set the elements of the matrix $\widehat{\mathcal{AC}}_{\lambda}$ corresponding to covariances between zero and nonzero elements to zero. It can be shown that this is equivalent to setting the corresponding entries of vectors $\text{vec}\mathbf{S}$ and $\text{vec}\mathbf{S}_k$ to zero. We define

$$\begin{aligned}\text{df}(\lambda) &:= \frac{1}{2n} \sum_{k=1}^n \left\{ \text{vec}(\mathbf{S}_k \circ \mathbf{I}_{\lambda})^{\top} \widehat{\mathcal{AC}}_{\lambda} \text{vec}(\mathbf{S}_k \circ \mathbf{I}_{\lambda}) \right\} \\ &\quad - \frac{1}{2} \left\{ \text{vec}(\mathbf{S} \circ \mathbf{I}_{\lambda})^{\top} \widehat{\mathcal{AC}}_{\lambda} (\mathbf{S} \circ \mathbf{I}_{\lambda}) \right\},\end{aligned}\quad (20)$$

where $\mathbf{I}_{\lambda}$ is the indicator matrix, whose entry is 1 if the corresponding entry in the precision matrix $\hat{\boldsymbol{\Omega}}_{\lambda}$ is nonzero and zero if the corresponding entry in the precision matrix is zero, and $\widehat{\mathcal{AC}}_{\lambda}$ is calculated by setting $\boldsymbol{\Omega} = \hat{\boldsymbol{\Omega}}_{\lambda}$ in (10).

Implementation of formula (20) is computationally expensive, so we apply Lemma 1 and the identity $\text{vec}(\mathbf{ABC}) = (\mathbf{C}^{\top} \otimes \mathbf{A})\text{vec}\mathbf{B}$ to (20) and obtain

$$\begin{aligned}\text{df}(\lambda) &= \frac{1}{2n} \sum_{k=1}^n \text{vec}(\mathbf{S}_k \circ \mathbf{I}_{\lambda})^{\top} \text{vec}\{\hat{\boldsymbol{\Omega}}_{\lambda}(\mathbf{S}_k \circ \mathbf{I}_{\lambda})\hat{\boldsymbol{\Omega}}_{\lambda}\} \\ &\quad - \frac{1}{2} \text{vec}(\mathbf{S} \circ \mathbf{I}_{\lambda})^{\top} \text{vec}\{\hat{\boldsymbol{\Omega}}_{\lambda}(\mathbf{S} \circ \mathbf{I}_{\lambda})\hat{\boldsymbol{\Omega}}_{\lambda}\}.\end{aligned}\quad (21)$$

For any pair of matrices $\mathbf{X} = (x_{ij})$ and $\mathbf{Y} = (y_{ij})$, the expression $(\text{vec}\mathbf{X})^{\top} \text{vec}\mathbf{Y}$ can be implemented efficiently with complexity $O(p^2)$.

Finally, the bias correction term in (21) can be used in model selection methods (3) and (4), so that the optimal λ would be their minimizer. Even though the formula (21) is based on the approximation and relies on the asymptotic covariance matrix of the MLE, the simulation study in Sect. 4 suggests that the proposed model selectors perform well.

4 Simulation study

In this section, first we compare the performance of several tuning parameter selectors for L_1 -norm penalized Gaussian graphical models. The tuning parameter λ is selected via GAIC or GBIC, and then the results are compared with respect to different model selection procedures. Specifically, we consider the estimator of degrees of freedom proposed in Yuan and Lin (2007)

$$\text{df}_{\text{Yuan}}(\lambda) := \sum_{1 \leq i < j \leq p} I(\omega_{ij,\lambda} \neq 0),\quad (22)$$

which uses (22), i.e., the number of nonzero elements, as bias term in AIC-type (3) and BIC-type (4) model selectors. A close competitor of the GAIC model selector is the Kullback–Leibler cross-validation (KLCV) model selector proposed in Vujačić et al. (2015). Secondly, we address the issue of whether the graphical lasso together with GAIC or GBIC is able to identify the true model consistently; finally, we compare the graphical lasso with the adaptive lasso and the SCAD in combination with GAIC and GBIC.

In all simulations, the solution path, which consists of the estimated precision matrices for different values of λ , is obtained by maximum penalized likelihood estimation as in Equation (2) using the graphical lasso algorithm Friedman et al. (2008), which builds on the block coordinate descent algorithm of d’Aspremont et al. (2008). To compute the solution path, we used `glassopath`, implemented in the R package `glasso`, with no penalization on the diagonal elements.

Tuning parameter selectors comparison. We generate data from multivariate Gaussian distributions $N(0, \Omega)$ where the precision Ω is associated with a sparse random graph. The sparsity of the graph depends on the probability that a link will be present which is fixed at $3/p$, where p is the number of nodes. The graphs have $p = 60, 120, 240$ nodes. The sample sizes, generated from the multivariate Gaussian distributions, are $n = 60$ and $n = 120$. We compare AIC, KLCV, and GAIC with respect to estimation quality (entropy loss), and BIC, KLCV, and GBIC with respect to support recovery of the precision matrix (F-score). The estimation quality is evaluated in terms of entropy loss (14). The smaller the entropy loss, the better the performance of the estimator in terms of estimation quality. Let TP be the true positives, FP be the false positives, TN be the true negatives, and FN be the false negatives. The support recovery of the precision matrix is evaluated in terms of F-score

$$\text{F-score} = \frac{2TP}{2TP + FP + FN}.$$

The F-score can be viewed as a weighted average of the precision and recall (F-score reaches its best value at 1 and worst score at 0).

Table 1 shows the results, in terms of averaged entropy loss over 100 simulated data, of the simulation study for $p = 60, 120, 240$, and $n = 60, 120$. This table suggests that AIC performs well when the number of variables is smaller or equal to the number of observations, whereas it has the worst performance otherwise. The KLCV and GAIC show similar performance. However, the latter reduces the computational burden (see Table S1 in the supplementary material). Table 2, which reports the mean F-score for BIC, KLCV, and GBIC, suggests that the BIC performs well when n is much

Table 1 Mean absolute entropy loss over 100 simulated data from multivariate Gaussian distribution $N(0, \Omega)$ for several model selections and scenarios

Entropy			
p	AIC	KLCV	GAIC
$n = 60$			
60	<u>3.16</u> (0.30)	3.31(0.26)	3.28(0.26)
120	359.01(45.37)	7.46(0.46)	<u>7.32</u> (0.37)
240	1668.74(180.43)	16.22(0.60)	<u>15.86</u> (0.61)
$n = 120$			
60	<u>1.69</u> (0.13)	1.79(0.13)	1.79(0.13)
120	<u>4.00</u> (0.27)	4.31(0.31)	4.30(0.31)
240	628.04(57.66)	10.68(0.44)	<u>10.63</u> (0.42)

A random graph structure was used to generate the precision matrix Ω . The precision matrix was estimated via graphical lasso approach. Standard errors are reported in brackets. The underlined elements indicate the best model selectors for each scenario

Table 2 Mean F-score of different estimators over 100 simulated data from multivariate Gaussian distribution $N(0, \Omega)$ for several model selections and scenarios

F-score			
p	BIC	KLCV	GBIC
$n = 60$			
60	0.30(0.22)	0.46(0.05)	<u>0.47</u> (0.05)
120	0.08(0.12)	0.41(0.03)	<u>0.42</u> (0.03)
240	0.00(0.01)	0.22(0.09)	<u>0.25</u> (0.05)
$n = 120$			
60	<u>0.72</u> (0.05)	0.48(0.05)	0.56(0.06)
120	<u>0.69</u> (0.05)	0.46(0.04)	0.52(0.04)
240	0.16(0.17)	0.40(0.03)	<u>0.48</u> (0.02)

A random graph structure was used to generate the precision matrix Ω . The precision matrix was estimated via graphical lasso approach. Standard errors are reported in brackets. The underlined elements indicate the best model selectors for each scenario

greater than p , whereas it has the worst performance for n smaller or even equal than p . KLCV and GAIC shows similar performance for n much greater than p , but the latter shows better performance otherwise.

We also compared the tuning parameter selectors changing the structure of the graph for the same combinations of n and p . Specifically, we considered dense random graph where the probability that a link will be present is fixed at $10/p$, cluster, and scale-free graphs. The results are shown in the supplementary material (Tables S2, S3, S4, S5, S6, and S7) and suggest no dependence on the graph structures.

Model consistency. In another simulation study, we address the issue of whether the graphical lasso with tuning parameter selected by one of the two proposed estimators (GAIC, GBIC) is able to identify the true model consistently and

Table 3 Mean Frobenius norm and mean F-score over 100 simulated data from multivariate Gaussian distribution $N(0, \Omega)$ for GAIC and GBIC model selections and scenarios

n	F-score		Frobenius norm	
	GAIC	GBIC	GAIC	GBIC
$p = 60$				
60	0.44(0.04)	0.44(0.04)	4.67(0.16)	4.69(0.16)
120	0.49(0.04)	0.55(0.05)	3.74(0.15)	3.93(0.18)
240	0.49(0.04)	0.61(0.04)	2.89(0.12)	3.19(0.13)
480	0.49(0.04)	0.63(0.04)	2.20(0.09)	2.49(0.12)
$p = 120$				
60	0.39(0.03)	0.40(0.03)	7.04(0.15)	7.08(0.18)
120	0.47(0.04)	0.54(0.04)	5.76(0.17)	6.03(0.16)
240	0.49(0.03)	0.61(0.05)	4.52(0.13)	4.84(0.15)
480	0.49(0.04)	0.66(0.04)	3.42(0.12)	3.82(0.11)

A random graph structure was used to generate the precision matrix Ω . The precision matrix was estimated via graphical lasso approach. The sample size n goes from 60 up to 480, and the number of variables p is set to 60 and 120. Standard errors are reported in brackets

whether it is capable of identifying the true graph. Model consistency of the proposed estimators is measured by the Frobenius norm $\|\Omega - \hat{\Omega}_\lambda\|_F$ and the capability of identifying the true graph is measured by the F-score. In this simulation scheme, we applied the graphical lasso algorithm to a random graph. The sample size n goes from 60 up to 480, and the number of variables p is set to 60 and 120.

Table 3 shows the mean Frobenius norm and the mean F-score over 100 simulated data. Both the estimators seem to be consistent in terms of Frobenius norm. In terms of F-score, GBIC approaches to 1, whereas GAIC does not show this property.

Comparison with other penalized models. In this simulation scheme, we compare the L_1 -norm penalty with the SCAD and the adaptive lasso penalties. The SCAD penalty is symmetric and a quadratic spline on $[0, \infty]$, whose first derivative is

$$\text{SCAD}'_{\lambda,a}(x) = \lambda \left\{ I(|x| \leq \lambda) + \frac{(a\lambda - |x|)_+}{(a-1)\lambda} I(|x| > \lambda) \right\}, \quad (23)$$

where $\lambda \geq 0$ and $a > 2$ are two tuning parameters. The adaptive lasso Zou (2006) is another method to achieve model selection and parameter estimation, simultaneously. Both these methods use weighted versions of the graphical lasso. For our setting, we define the SCAD weights $w_{ij} = \text{SCAD}'_{\lambda,a}(|\tilde{w}_{ij}|)$ with $a = 3.7$ and the adaptive weights to be $w_{ij} = 1/|\tilde{w}_{ij}|^\gamma$ with $\gamma = 0.5$. We generate data from multivariate Gaussian distributions $N(0, \Omega)$ where the precision Ω gives a sparse random graph. The probability that a link is

Table 4 Mean absolute Frobenius norm over 100 simulated data from multivariate Gaussian distribution $N(0, \Omega)$ for GAIC and GBIC model selections and scenarios

p	Frobenius		
	Glasso	Adaptive	SCAD
$n = 60$			
60	4.73(0.18)	<u>4.45</u> (0.24)	4.72(0.20)
120	6.78(0.18)	<u>6.71</u> (0.30)	6.77(0.18)
240	<u>8.76</u> (0.18)	9.53(0.47)	8.77(0.18)
$n = 120$			
60	3.87(0.16)	<u>3.17</u> (0.20)	3.71(0.18)
120	5.53(0.12)	<u>4.87</u> (0.24)	5.47(0.14)
240	<u>7.37</u> (0.15)	7.72(0.26)	7.38(0.15)

A random graph structure was used to generate the precision matrix Ω . The precision matrix was estimated via graphical lasso, adaptive lasso, and SCAD. For each weighted algorithm, the optimum value of the tuning parameter λ was selected by GAIC. Standard errors are reported in brackets. The underlined elements indicate the best model selectors for each scenario

Table 5 Mean F-score of different estimators over 100 simulated data from multivariate Gaussian distribution $N(0, \Omega)$ for GAIC and GBIC model selections and scenarios. A random graph structure was used to generate the precision matrix Ω . The precision matrix was estimated via graphical lasso, adaptive lasso, and SCAD. For each weighted algorithm, the optimum value of the tuning parameter λ was selected by GBIC. Standard errors are reported in brackets. The underlined elements indicate the best model selectors for each scenario

p	F-score		
	Glasso	Adaptive	SCAD
$n = 60$			
60	0.44(0.03)	<u>0.48</u> (0.04)	0.44(0.03)
120	0.39(0.03)	<u>0.41</u> (0.03)	0.39(0.03)
240	0.25(0.05)	0.25(0.05)	0.25(0.05)
$n = 120$			
60	0.55(0.04)	<u>0.63</u> (0.04)	0.56(0.05)
120	0.52(0.03)	<u>0.56</u> (0.03)	0.53(0.03)
240	0.47(0.03)	<u>0.48</u> (0.03)	0.46(0.03)

present is fixed at $3/p$, where p is the number of variables. The graphs have $p = 60, 120, 240$ nodes. The sample sizes, generated from the multivariate Gaussian distributions, are $n = 60$ and $n = 120$. We compare glasso, SCAD, and adaptive lasso in terms of estimation quality (Frobenius norm) and support recovery of the precision matrix (F-score). Tables 4 and 5 show the mean absolute Frobenius norm and the mean F-score over 100 simulated data for graphical lasso, adaptive lasso, and SCAD, respectively. The adaptive lasso improves the quality of the estimator both in terms of Frobenius norm and F-score. Instead, the SCAD penalty does not show any improvement with respect to the L_1 -norm.

Fig. 1 Tuning parameter selection criteria applied to Arabidopsis dataset in which we considered 118 samples and 39 genes

5 Application of GLC to biological network reconstruction

In this section, we test the robustness of the method on a genomic dataset. In particular, we consider gene expression data from Wille et al. (2004) on the extensively studied *Arabidopsis thaliana* biological system. The study reports data from $n = 118$ experiments on $p = 39$ genes, whose regulatory network is of interest. Figure 1 shows AIC, BIC, GAIC, and GBIC (y-axis) for a set of λ (x-axis) resulting from the application of the glasso algorithm to the dataset Arabidopsis. Note that we have omitted KLCV from the figure since it shows a curve similar to the GAIC. The “best” values of λ according to both AIC and BIC are very close to zero, which means that the resulting AIC and BIC graphs will be quite dense, whereas GAIC, KLCV, and GBIC seem to give more sensible results. In order to select more robust

graphs, we used a re-sample technique. In particular, we re-sampled 100 times without replicates 70% of the samples and for each of this re-sampled data we computed the best graph according to AIC, BIC, KLCV, GAIC, GBIC. Then, for each of these information criteria we computed the averaged adjacent matrices, which gives a score (between 0 and 1) for each link of the graph. The GBIC, KLCV, and BIC graphs built by selecting only the edges with scores greater than 0.7 are shown in Fig. 2. Note that both the GBIC and KLCV graphs are remarkably simple and informative, exhibiting some cliques and hub genes. The GBIC graph is slightly sparser than the KLCV graph. In contrast, the BIC graph is very dense and possible useful association information is buried in the large number of estimated edges. The most connected nodes are MPDC1, HDS, MK DXR, and HMGR2. The degree distribution table is shown in Supplementary Materials (Table S8).

6 Discussion

High-dimensional data arise in many fields of science. The study of the structure, in terms of the conditional independence graph, can be carried out with penalized Gaussian graphical models. These models allow us to visualize with a graph the conditional independence among the set of random variables. These models obtain sparse graphs by means of heavy penalization, i.e., many parameters are estimated as zero. In this setting, we have proposed an estimator of the Kullback–Leibler loss function which can be used for selecting the tuning parameter in the main penalized likelihood approach such as glasso, adaptive lasso, and SCAD. In a simulation study, we evaluated the performance of the proposed estimator and compared it with others tuning

Fig. 2 Gene expression study. The estimated conditional independence graphs are the results of the application of the graphical lasso algorithm to the dataset *Arabidopsis thaliana*. The precision matrix was estimated by selecting the best tuning parameter according to GBIC, KLCV, and

BIC. The graphs derived from GBIC is shown on the left side, the one derived from KLCV is in the center, and the one derived from BIC is visualized on the right side. The degree for each node is reported in Supplementary Materials Table S8

parameter selectors. Our estimator can be combined with information criteria to improve their performance in terms of both estimation quality and support recovery of the precision matrix. Although the GAIC minimizes the KL divergence, we have shown that GBIC can also be useful for graph selection.

Acknowledgements The project was partially supported by the “European Cooperation in Science & Technology” (COST) funding: action number CA15109.

A Appendix

Lemma 1 Let $\mathbf{A}, \mathbf{B}$ be symmetric matrices of order p and $\mathbf{\Omega}$ be a positive definite symmetric matrix of order p . Then, the following identity holds

$$\langle \text{vec} \mathbf{A}, \text{vec} \mathbf{B} \rangle_{\mathbf{\Omega} \otimes \mathbf{\Omega}} = \langle \text{vec} \mathbf{A}, \text{vec} \mathbf{B} \rangle_{\mathbf{M}_p(\mathbf{\Omega} \otimes \mathbf{\Omega})},$$

where $\langle \text{vec} \mathbf{A}, \text{vec} \mathbf{B} \rangle_{\mathbf{\Psi}} = \text{vec} \mathbf{B}^T \mathbf{\Psi} \text{vec} \mathbf{A}$ is the inner product induced by positive definite $p^2 \times p^2$ matrix $\mathbf{\Psi}$, and $\mathbf{M}_p = \frac{1}{2}(\mathbf{I}_{p^2} + \mathbf{K}_p)$.

Proof The lemma follows from the definition of the inner product and the identity $(\mathbf{\Omega} \otimes \mathbf{\Omega}) \text{vec} \mathbf{A} = \mathbf{M}_p(\mathbf{\Omega} \otimes \mathbf{\Omega}) \text{vec} \mathbf{A}$ which is proved in Vujčić et al. (2015), Lemma 4.1. $\square$

Before we prove Lemma 2, we prove some auxiliary results.

Lemma A.1 Let $\mathbf{A} = (a_{ij})$ be an $n \times n$ matrix, $I = \{i_1, \dots, i_s\}$ the subset of $\{1, \dots, n\}$, π any permutation of the set $\{1, \dots, n\}$ and $\mathbf{X} = (x_{ij})$ an $n \times n$ matrix such that

$$x_{ij} = \begin{cases} x, & i \in I; j = \pi(i) \\ 0, & \text{otherwise} \end{cases},$$

where $x \in \mathbb{R}$. Denote by $j_k = \pi(i_k)$, $k = 1, \dots, s$, and $J = \{j_1, \dots, j_s\}$, then

$$|\mathbf{A} + \mathbf{X}| = (-1)^l |\mathbf{A}[I^C, J^C]| x^s + p_{s-1}(x),$$

where $l = \sum_{k=1}^s (i_k + j_k)$ and $p_{s-1}(x)$ is the polynomial in x of degree $s - 1$.

Proof Laplace expansion and induction on s yield the stated result. $\square$

Roughly speaking, the lemma states that if an element x is appearing s times in different rows and columns of a matrix, then the determinant of that matrix is a polynomial in x of degree s .

Lemma 2 Let $I_\lambda(\alpha) \subset \{1, \dots, p^2\}$ be the set of indices of nonzero entries in the diagonal of $\mathbf{D}_\lambda(\alpha)$ and I_λ be the set of indices of nonzero entries in $\text{vec} \hat{\mathbf{\Omega}}_\lambda$.

1. If there exists α_0 such that $I_\lambda(\alpha) = \{1, \dots, p^2\}$, for $\alpha \geq \alpha_0$, then the limit $\lim_{\alpha \rightarrow +\infty} \hat{\mathcal{I}}_\lambda(\alpha)^{-1}$, denoted by $\hat{\mathcal{I}}_\lambda^{-1}$, exists and

$$\hat{\mathcal{I}}_\lambda^{-1} = \hat{\mathbf{\Omega}}_\lambda \otimes \hat{\mathbf{\Omega}}_\lambda. \quad (18)$$

2. If there exists α_0 such that $I_\lambda(\alpha) = I_\lambda$ for $\alpha \geq \alpha_0$, then the limit $\lim_{\alpha \rightarrow +\infty} \hat{\mathcal{I}}_\lambda(\alpha)^{-1}$, denoted by $\hat{\mathcal{I}}_\lambda^{-1}$, exists and

$$\begin{aligned} (\hat{\mathcal{I}}_\lambda^{-1})_{i,j} &= 0, \quad \text{when } i \text{ or } j \notin I_\lambda, \\ \hat{\mathcal{I}}_\lambda^{-1}[I_\lambda, I_\lambda] &= \left\{ (\hat{\mathbf{\Omega}}_\lambda^{-1} \otimes \hat{\mathbf{\Omega}}_\lambda^{-1})[I_\lambda, I_\lambda] \right\}^{-1}. \end{aligned} \quad (19)$$

Proof With the notation $\mathbf{A}_\lambda(\alpha) = \hat{\mathbf{\Omega}}_{\lambda,\alpha}^{-1} \otimes \hat{\mathbf{\Omega}}_{\lambda,\alpha}^{-1}$, we have that $\hat{\mathcal{I}}_\lambda(\alpha) = \mathbf{A}_\lambda(\alpha) + \mathbf{D}_\lambda(\alpha)$, where $\mathbf{D}(\alpha)$ is defined after formula (17). For the function defined in (16), we have:

$$\mathbf{R}_\lambda(\alpha) = \mathbf{A}_\lambda(\alpha) + \lambda \mathbf{D}_\lambda(\alpha). \quad (24)$$

Here and subsequently, $o(1)$ stands for the infinitely small function when $\alpha \rightarrow \infty$, and all limits will be with respect to this base. For the approximated problem defined by function (16), it follows that a diagonal element in $\mathbf{D}_\lambda(\alpha)$ is equal to $\alpha/2$ if the corresponding $\hat{\omega}_{\lambda,ij}$ is zero and equal to $o(1)$ if it is nonzero.

1. According to the assumption of the lemma, we have $I_\lambda(\alpha) = \{1, \dots, p^2\}$, and therefore all the diagonal elements of $\mathbf{D}_\lambda(\alpha)$ are equal to $o(1)$. The result now follows from continuity of the matrix inversion.
2. According to the assumption of the lemma, there exists α_0 such that $I(\alpha) = I_\lambda$ for $\alpha \geq \alpha_0$. Fix $\alpha \geq \alpha_0$. To show (19), we will use the Kramer’s rule.

Fix the (i, j) th element of $\hat{\mathcal{I}}_\lambda^{-1}(\alpha)$. If i or j do not belong to the set I_λ , then by Kramer’s rule and Lemma A.2 (i), (ii.a)

$$\begin{aligned} (\hat{\mathcal{I}}_\lambda^{-1}(\alpha))_{ij} &= \frac{C_{ij}\{\hat{\mathcal{I}}_\lambda(\alpha)\}}{|\hat{\mathcal{I}}_\lambda(\alpha)|} \\ &= \frac{q_{s-1}(\alpha/2)}{(\alpha/2)^s |\hat{\mathcal{I}}_\lambda(\alpha)[I_\lambda, I_\lambda]| + p(\alpha/2)} \\ &= o(1/\alpha). \end{aligned} \quad (25)$$

For the case when $i, j \in I_\lambda$, we use the fact that

$$\lim_{\alpha \rightarrow +\infty} \hat{\mathcal{I}}_\lambda(\alpha)[I, I] = \mathbf{A}_\lambda(\alpha)[I_\lambda, I_\lambda],$$

since $\mathbf{A}_\lambda(\alpha)[I_\lambda, I_\lambda]$ and $\hat{\mathbf{T}}_\lambda(\alpha)[I_\lambda, I_\lambda]$ contain the same elements, except on the diagonal, where they differ in $o(1)$ term. Since determinant is continuous function it follows that $C_{\tilde{i}, \tilde{j}}\{\hat{\mathbf{T}}_\lambda(\alpha)[I_\lambda, I_\lambda]\} = C_{\tilde{i}, \tilde{j}}\{\mathbf{A}_\lambda(\alpha)[I_\lambda, I_\lambda]\} + o(1)$ and $|\hat{\mathbf{T}}_\lambda(\alpha)[I_\lambda, I_\lambda]| = |\mathbf{A}_\lambda(\alpha)[I_\lambda, I_\lambda]| + o(1)$. Using Kramer's rule, Lemma A.2 (i), (ii.b), and the results shown above, we obtain

$$\begin{aligned} (\hat{\mathbf{T}}_\lambda^{-1}(\alpha))_{ij} &= \frac{C_{ij}\{\hat{\mathbf{T}}_\lambda(\alpha)\}}{|\hat{\mathbf{T}}_\lambda(\alpha)|} \\ &= \frac{(\alpha/2)^s C_{\tilde{i}, \tilde{j}}\{\hat{\mathbf{T}}_\lambda(\alpha)[I_\lambda, I_\lambda]\} + q_{s-1}(\alpha/2)}{(\alpha/2)^s |\hat{\mathbf{T}}_\lambda(\alpha)[I_\lambda, I_\lambda]| + p_{s-1}(\alpha/2)} \\ &= \frac{C_{\tilde{i}, \tilde{j}}\{\mathbf{A}_\lambda(\alpha)[I_\lambda, I_\lambda]\} + o(1)}{|\mathbf{A}_\lambda(\alpha)[I_\lambda, I_\lambda]| + o(1)}. \end{aligned} \quad (26)$$

Taking the limit in (25) and (26), we obtain (19). $\square$

The proof of Lemma 2 holds for any approximation of the penalty function whose second derivative does not depend on Ω but only on α .

Lemma A.2 *The following holds*

- (i) $|\hat{\mathbf{T}}_\lambda(\alpha)| = (\alpha/2)^s |\hat{\mathbf{T}}_\lambda(\alpha)[I_\lambda, I_\lambda]| + p_{s-1}(\alpha/2)$, where $p_{s-1}(\alpha/2)$ is a polynomial in $\alpha/2$ of degree $s-1$ and s is the number of zero elements in the matrix $\hat{\Omega}_\lambda$.
- (ii.a) If i or j do not belong to the set I_λ , then $C_{i,j}\{\hat{\mathbf{T}}_\lambda(\alpha)\}$ is a polynomial in $\alpha/2$ of degree $s-1$.
- (ii.b) If i, j are elements of I_λ , then $C_{i,j}\{\mathbf{T}_\lambda(\alpha)\} = (\alpha/2)^s C_{\tilde{i}, \tilde{j}}\{\hat{\mathbf{T}}_\lambda(\alpha)[I_\lambda, I_\lambda]\} + q_{s-1}(\alpha/2)$, where q is a polynomial in $\alpha/2$ of degree $s-1$ and $(\tilde{i}, \tilde{j})$ is position of element $(\hat{\mathbf{T}}_\lambda(\alpha))_{i,j}$ in the matrix $\hat{\mathbf{T}}_\lambda(\alpha)[I_\lambda, I_\lambda]$.

Proof Assertions (i) and (ii.a) follow by applying Lemma A.1.

To prove (ii.b) assume that $i, j \in I_\lambda$. Throughout the proof, we refer to element of the matrix that contain α as α -element, and the row (column) that contain α -element as α -row (column).

Denote the number of α -rows with index less than i as r and number of α -columns with index less than j as c . It is straightforward to check that the position of element $(\hat{\mathbf{T}}_\lambda(\alpha))_{ij}$ in the matrix $\hat{\mathbf{T}}_\lambda(\alpha)[I_\lambda, I_\lambda]$ is $(i-r, j-c)$, i.e., $\tilde{i} = i-r, \tilde{j} = j-c$. It follows that (ii.b) is equivalent to

$$\begin{aligned} (-1)^{i+j} M_{i,j}\{\hat{\mathbf{T}}_\lambda(\alpha)\} &= (-1)^{i+j-r-c} (\alpha/2)^s \\ &\quad M_{\tilde{i}, \tilde{j}}\{\hat{\mathbf{T}}_\lambda(\alpha)[I_\lambda, I_\lambda]\} + q_{s-1}(\alpha/2). \end{aligned}$$

To show this, we consider three cases.

- $i = j$. Applying Lemma A.1 to $M_{i,j}\{\hat{\mathbf{T}}_\lambda(\alpha)\}$ and using $l = \sum_{k=1}^s (i_k + j_k) = 2 \sum_{k=1}^s i_k$, we obtain

$$\begin{aligned} M_{i,j}\{\hat{\mathbf{T}}_\lambda(\alpha)\} &= (-1)^l (\alpha/2)^s M_{\tilde{i}, \tilde{j}}\{\hat{\mathbf{T}}_\lambda(\alpha)[I_\lambda, I_\lambda]\} \\ &\quad + q_{s-1}(\alpha/2) \\ &= (\alpha/2)^s M_{\tilde{i}, \tilde{j}}\{\hat{\mathbf{T}}_\lambda(\alpha)[I_\lambda, I_\lambda]\} + q_{s-1}(\alpha/2), \end{aligned}$$

In this case, $r = c$ so $(-1)^{i-r+j-c} = (-1)^{i+j-2r} = (-1)^{i+j}$. From this, the equality of cofactors follows.

- $i < j$. In this case, of all α -elements $c-r$ of them are not in the diagonal of the matrix $\mathbf{A}_\lambda(\alpha)[\{i\}^C, \{j\}^C]$, but on the superdiagonal. Hence, by applying Lemma A.1 to $M_{i,j}\{\hat{\mathbf{T}}_\lambda(\alpha)\}$ we obtain $l = \sum_{k=1}^s (i_k + j_k) = c-r$ and consequently

$$\begin{aligned} M_{i,j}\{\hat{\mathbf{T}}_\lambda(\alpha)\} &= (-1)^l (\alpha/2)^s M_{\tilde{i}, \tilde{j}}\{\hat{\mathbf{T}}_\lambda(\alpha)[I_\lambda, I_\lambda]\} \\ &\quad + q_{s-1}(\alpha/2) \\ &= (-1)^{c-r} (\alpha/2)^s M_{\tilde{i}, \tilde{j}}\{\hat{\mathbf{T}}_\lambda(\alpha)[I_\lambda, I_\lambda]\} \\ &\quad + q_{s-1}(\alpha/2). \end{aligned}$$

Since $(-1)^{c-r} = (-1)^{-r-c+2c} = (-1)^{-r-c}$, multiplying both sides of the last equality by $(-1)^{i+j}$ yields the equality of cofactors.

- $i > j$. Analogue to the previous case, here we use that of all α -elements $r-c$ of them are on the subdiagonal and equality $(-1)^{r-c} = (-1)^{-r-c}$. $\square$

References

- Bogdan, M., Ghosh, J.K., Doerge, R.W.: Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci. *Genetics* **167**(2), 989–999 (2004)
- Broman, K.W., Speed, T.P.: A model selection approach for the identification of quantitative trait loci in experimental crosses. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **64**(4), 641–656 (2002)
- Cai, T., Liu, W., Luo, X.: A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Am. Stat. Assoc.* **106**(3), 594–607 (2011)
- Craven, P., Wahba, G.: Smoothing noisy data with spline functions. *Numer. Math.* **31**(4), 377–403 (1978)
- Chen, J., Chen, Z.: Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95**(3), 759–771 (2008)
- d'Aspremont, A., Banerjee, O., El Ghaoui, L.: First-order methods for sparse covariance selection. *SIAM J. Matrix Anal. Appl.* **30**(1), 56–66 (2008)
- Dong, X., Wahba, G.: A generalized approximate cross validation for smoothing splines with non-Gaussian data. *Stat. Sin.* **6**(3), 675–692 (1996)
- Edwards, D.: *Introduction to Graphical Modelling*. Springer, Berlin (2000)
- Fan, J., Feng, Y., Wu, Y.: Network exploration via the adaptive lasso and scad penalties. *Ann. Appl. Stat.* **3**(2), 521–541 (2009)
- Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **96**(456), 1348–1360 (2001)

- Foygel, R., Drton, M.: Extended bayesian information criteria for gaussian graphical models. *Adv. Neural Inf. Process. Syst.* **23**, 604–612 (2010)
- Fried, R., Vogel, D.: On robust gaussian Graphical Modelling. *Recent Developments in Applied Probability and Statistics*. Springer, Berlin (2009)
- Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**(3), 432–441 (2008)
- Gao, X., Pu, D.Q., Wu, Y., Xu, H.: Tuning parameter selection for penalized likelihood estimation of Gaussian graphical model. *Stat. Sin.* **22**(3), 1123 (2012)
- Konishi, S., Kitagawa, G.: Generalised information criteria in model selection. *Biometrika* **83**(4), 875–890 (1996)
- Konishi, S., Kitagawa, G.: *Information Criteria and Statistical Modeling*. Springer, Berlin (2008)
- Lam, C., Fan, J.: Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Stat.* **37**(6B), 4254 (2009)
- Lauritzen, S.L.: *Graphical Models*. Oxford University Press, Oxford (1996)
- Lian, H.: Shrinkage tuning parameter selection in precision matrices estimation. *J. Stat. Plan. Inference* **141**(8), 2839–2848 (2011)
- Liu, H., Roeder, K., Wasserman, L.: Stability approach to regularization selection (stars) for high dimensional graphical models. *Adv. Neural Inf. Process. Syst.* **23**, 1432–1440 (2010)
- Magnus, J.R., Neudecker, H.: *Matrix Differential Calculus with Application in Statistics and Econometrics*. Wiley Series in Probability and Statistics. Wiley, Hoboken (2007)
- Nishii, R.: Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Stat.* **12**(2), 758765 (1984)
- Penny, W.: Kullback–Leibler divergences of normal, gamma, Dirichlet and Wishart densities. *Tech. Rep.*, Wellcome Department of Cognitive Neurology (2001)
- Rothman, A.J., Bickel, P.J., Levina, E., Zhu, J.: Sparse permutation invariant covariance estimation. *Electr. J. Stat.* **2**, 494–515 (2008)
- Rue, H., Held, L.: *Gaussian Markov Random Fields: Theory and Applications*. CRC Press, Boca Raton (2005)
- Schmidt, M., Fung, G. and Rosales, R.: Fast optimization methods for l1 regularization: A comparative study and two new approaches. In: *Proceedings of European Conference on Machine Learning*, pp. 286–297. Springer (2007)
- Siegmund, D.: Model selection in irregular problems: applications to mapping quantitative trait loci. *Biometrika* **94**(3), 553–568 (2004)
- Zhang, Y., Li, R., Tsai, C.-L.: Regularization parameter selections via generalized information criterion. *J. Am. Stat. Assoc.* **105**(489), 312323 (2010)
- Wang, H., Li, R., Tsai, C.-L.: Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **91**(4), 785–800 (2007)
- Zou, H.: The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* **101**(476), 1418–1429 (2006)
- Whittaker, J.: *Graphical Models in Applied Multivariate Statistics*. Wiley Publishing, Hoboken (2009)
- Yuan, M., Lin, Y.: Model selection and estimation in the Gaussian graphical model. *Biometrika* **94**(1), 19–35 (2007)
- Zhao, T., Liu, H., Roeder, K., Lafferty, J., Wasserman, L.: The huge package for high-dimensional undirected graph estimation in R. *J. Mach. Learn. Res.* **13**(4), 1059–1062 (2012)
- Vujačić, I., Abbruzzo, A., Wit, E.: A computationally fast alternative to cross-validation in penalized Gaussian graphical models. *J. Stat. Comput. Simul.* **85**(18), 3628–3640 (2015)
- Wille, A., Zimmermann, P., Vranová, E., Fürholz, A., Laule, O., Bleuler, S., Hennig, L., Prelić, A., von Rohr, P., Thiele, L., Zitzler, E., Gruissem, W., Buhlmann, P.: Sparse graphical Gaussian modeling of the isoprenoid gene network in *Arabidopsis thaliana*. *Genome Biol.* **5**(11), R92 (2004)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.