Statistics and Computing (2020) 30:351–376
https://doi.org/10.1007/s11222-019-09879-9

Robust Bayesian model selection for variable clustering with the Gaussian graphical model

Daniel Andrade[1,2] · **Akiko Takeda**[3,4] · **Kenji Fukumizu**[5]

Received: 15 June 2018 / Accepted: 1 June 2019 / Published online: 19 July 2019

Abstract
Variable clustering is important for explanatory analysis. However, only few dedicated methods for variable clustering with the Gaussian graphical model have been proposed. Even more severe, small insignificant partial correlations due to noise can dramatically change the clustering result when evaluating for example with the Bayesian information criteria (BIC). In this work, we try to address this issue by proposing a Bayesian model that accounts for negligible small, but not necessarily zero, partial correlations. Based on our model, we propose to evaluate a variable clustering result using the marginal likelihood. To address the intractable calculation of the marginal likelihood, we propose two solutions: one based on a variational approximation and another based on MCMC. Experiments on simulated data show that the proposed method is similarly accurate as BIC in the no noise setting, but considerably more accurate when there are noisy partial correlations. Furthermore, on real data the proposed method provides clustering results that are intuitively sensible, which is not always the case when using BIC or its extensions.

Keywords Clustering · Gaussian graphical model · Model selection · Variational approximation

1 Introduction

The Gaussian graphical model (GGM) has become an invaluable tool for detecting partial correlations between variables. Assuming the variables are jointly drawn from a multivariate normal distribution, the sparsity pattern of the precision matrix reveals which pairs of variables are independent given all other variables (Anderson 2004). In particular, we can find clusters of variables that are mutually independent, by grouping the variables according their entries in the precision matrix.

For example, in gene expression analysis, variable clustering is often considered to be helpful for data exploration (Palla et al. 2012; Tan et al. 2015).

However, in practice, it can be difficult to find a meaningful clustering due to the noise of the entries in the partial correlations. The noise can be due to the sampling, this is in particular the case when n the number of observations is small, or due to small nonzero partial correlations in the true precision matrix that might be considered as insignificant. Here in this work, we are particularly interested in the latter type of noise. In the extreme, small partial correlations might lead to a connected graph of variables, where no grouping of variables can be identified. For an exploratory analysis, such a result might not be desirable.

As an alternative, we propose to cluster variables, such that the partial correlation between any two variables in different clusters is negligibly small, but not necessarily zero. The open question, which we try to address here, is whether there is a principled model selection criteria for this scenario.

✉ Daniel Andrade
andrade@ism.ac.jp

Akiko Takeda
takeda@mist.i.u-tokyo.ac.jp

Kenji Fukumizu
fukumizu@ism.ac.jp

[1] SOKENDAI (The Graduate University for Advanced Studies), 10-3 Midoricho, Tachikawa, Tokyo 190-8562, Japan

[2] Security Research Laboratories, NEC, 1753, Shimonumabe, Nakahara-ku, Kawasaki 211-8666, Japan

[3] Department of Creative Informatics, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

[4] RIKEN Center for Advanced Intelligence Project, 1-4-1, Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan

[5] The Institute of Statistical Mathematics, 10-3 Midoricho, Tachikawa, Tokyo 190-8562, Japan

For example, the Bayesian information criterion (BIC) (Schwarz 1978) is a popular model selection criterion for the Gaussian graphical model. However, in the noise setting it does not have any formal guarantees. As a solution, we propose here a Bayesian model that explicitly accounts for small partial correlations between variables in different clusters.

Under our proposed model, the marginal likelihood of the data can then be used to identify the correct (if there is a ground truth in theory), or at least a meaningful clustering (in practice) that helps analysis. Since the marginal likelihood of our model does not have an analytic solution, we provide two approximations: the first is a variational approximation, and the second is based on MCMC.

Experiments on simulated data show that the proposed method is similarly accurate as BIC in the no noise setting, but considerably more accurate when there are noisy partial correlations. The proposed method also compares favorable to two previously proposed methods for variable clustering and model selection, namely the Clustered Graphical Lasso (CGL) (Tan et al. 2015) and the Dirichlet Process Variable Clustering (DPVC) (Palla et al. 2012) method.

Our paper is organized as follows. In Sect. 2, we discuss previous works related to variable clustering and model selection. In Sect. 3, we introduce a basic Bayesian model for evaluating variable clusterings, which we then extend in Sect. 4.1 to handle noise on the precision matrix. For the proposed model, the calculation of the marginal likelihood is infeasible and we describe two approximation strategies in Sect. 4.2. Furthermore, since enumerating all possible clusterings is also intractable, we describe in Sect. 4.3 an heuristic based on spectral clustering to limit the number of candidate clusterings. We evaluate the proposed method on synthetic and real data in Sects. 5 and 6, respectively. Finally, we discuss our findings in Sect. 7.

2 Related work

Finding a clustering of variables is equivalent to finding an appropriate block structure of the covariance matrix. Recently, Tan et al. (2015) and Devijver and Gallopin (2018) suggested to detect block diagonal structure by thresholding the absolute values of the covariance matrix. Their methods perform model selection using the mean squared error of randomly left-out elements of the covariance matrix (Tan et al. 2015), and a slope heuristic (Devijver and Gallopin 2018).

Also several Bayesian latent variable models have been proposed for this task (Marlin and Murphy 2009; Sun et al. 2014; Palla et al. 2012). Each clustering, including the number of clusters, is either evaluated using the variational lower bound (Marlin and Murphy 2009), or by placing a Dirichlet process prior over clusterings (Palla et al. 2012; Sun et al. 2014). However, all of the above methods assume that the partial correlations of variables across clusters are exactly zero.

An exception is the work in Marlin et al. (2009) which proposes to regularize the precision matrix such that partial correlations of variables that belong to the same cluster are penalized less than those belonging to different clusters. For that purpose they introduce three hyper-parameters, λ_1 (for within-cluster penalty), λ_0 (for across clusters), with $\lambda_0 > \lambda_1$, and λ_D for a penalty of the diagonal elements. The clusters do not need to be known a priori and are estimated by optimizing a lower bound on the marginal likelihood. As such their method can also find variable clusterings, even when the true partial correlation of variables in different clusters is not exactly zero. However, the clustering result is influenced by three hyper-parameters λ_0, λ_1, and λ_D which have to be determined using cross-validation.

Recently, the work in Sun et al. (2015) and Hosseini and Lee (2016) relaxes the assumption of a clean block structure by allowing some variables to correspond to two clusters. The model selection issue, in particular, determining the number of clusters, is either addressed with some heuristics (Sun et al. 2015) or cross-validation (Hosseini and Lee 2016).

3 The Bayesian Gaussian graphical model for clustering

Our starting point for variable clustering is the following Bayesian Gaussian graphical model. Let us denote by d the number of variables, and n the number of observations. We assume that each observation $\mathbf{x} \in \mathbb{R}^d$ is generated i.i.d. from a multivariate normal distribution with zero mean and covariance matrix Σ. Assuming that there are k groups of variables that are mutually independent, we know that, after appropriate permutation of the variables, Σ has the following block structure

$$\Sigma = \begin{pmatrix} \Sigma_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \Sigma_k \end{pmatrix},$$

where $\Sigma_j \in \mathbb{R}^{d_j \times d_j}$, and d_j is the number of variables in cluster j.

By placing an inverse Wishart prior over each block Σ_j, we arrive at the following Bayesian model

$$\begin{aligned} &p(\mathbf{x}_1, \ldots, \mathbf{x}_n, \Sigma | \{\nu_j\}_j, \{\Sigma_{j,0}\}_j, \mathcal{C}) \\ &= \prod_{i=1}^{n} \text{Normal}(\mathbf{x}_i | \mathbf{0}, \Sigma) \prod_{j=1}^{k} \text{InvW}(\Sigma_j | \nu_j, \Sigma_{j,0}), \end{aligned} \tag{1}$$

where ν_j and $\Sigma_{j,0}$, are the degrees of freedom and the scale matrix, respectively. We set $\nu_j = d_j + 1$, $\Sigma_j = I_{d_j}$ lead-

ing to a non-informative prior on Σ_j. $\mathcal{C}$ denotes the variable clustering which imposes the block structure on Σ. We will refer to this model as the basic inverse Wishart prior model.

Assuming we are given a set of possible variable clusterings $\mathscr{C}$, we can then choose the clustering $\hat{\mathcal{C}}$ that maximizes the posterior probability of the clustering, i.e.,

$$\hat{\mathcal{C}} = \underset{\mathcal{C}\in\mathscr{C}}{\arg\max}\, p(\mathcal{C}|\mathscr{X}) = \underset{\mathcal{C}\in\mathscr{C}}{\arg\max}\, p(\mathscr{X}|\mathcal{C}) \cdot p(\mathcal{C}), \tag{2}$$

where we denote by $\mathscr{X}$ the observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$, and $p(\mathcal{C})$ is a prior over the clusterings which we assume to be uniform. Here, we refer to $p(\mathscr{X}|\mathcal{C})$ as the marginal likelihood (given the clustering). For the basic inverse Wishart prior model, the marginal likelihood can be calculated analytically, see, e.g., (Lenkoski and Dobra 2011).

4 Proposed method

In this section, we introduce our proposed method for finding variable clusters.

First, in Sect. 4.1, we extend the basic inverse Wishart prior model from Eq. (1) in order to account for nonzero partial correlations between variables in different clusters. Given the proposed model, the marginal likelihood $p(\mathscr{X}|\mathcal{C})$ does not have a closed form solution anymore. Therefore, in Sects. 4.2.2 and 4.2.3, we discuss two different methods for approximating the marginal likelihood. The first method is based on a variational approximation around the maximum a posteriori (MAP) solution. The second method is an MCMC method based on Chib's method (Chib 1995; Chib and Jeliazkov 2001). The latter has the advantage of being asymptotically correct for large number of posterior samples, but at considerably high computational costs. The former is considerably faster to evaluate and experimentally produces solutions similar to the MCMC method (see comparison in Sect. 5.3).

Finally, in Sect. 4.3, we propose to use a spectral clustering method to limit the clustering candidates to a set $\mathscr{C}^*$, where $\mathscr{C}^* \subseteq \mathscr{C}$. Based on this subset $\mathscr{C}^*$, we can then select the model maximizing the posterior probability [as in Eq. (2)], or can also calculate approximate posterior distributions over clusterings. We restrict the hypotheses space to $\mathscr{C}^*$, since even for a moderate number of variables, say $d = 40$, the size of the hypotheses space $|\mathscr{C}|$ is $> 10^{36}$. Therefore, MCMC sampling over the hypotheses space could also only explore a small subset of the whole hypotheses space, but at higher computational costs [see also Hans et al. (2007), Scott and Carvalho (2008) for a discussion on related high-dimensional problems].

4.1 A Bayesian Gaussian graphical model for clustering under noisy conditions

In this section, we extend the Bayesian model from Eq. (1) to account for nonzero partial correlations between variables in different clusters. For that purpose, we introduce the matrix $\Sigma_\epsilon \in \mathbb{R}^{d\times d}$ that models the noise on the precision matrix. The full joint probability of our model is given as follows:

$$\begin{aligned} &p(\mathbf{x}_1, \ldots, \mathbf{x}_n, \Sigma, \Sigma_\epsilon | \nu_\epsilon, \Sigma_{\epsilon,0}, \{\nu_j\}_j, \{\Sigma_{j,0}\}_j, \mathcal{C}) \\ &= \prod_{i=1}^{n} \text{Normal}(\mathbf{x}_i|\mathbf{0}, \Xi) \\ &\cdot \text{InvW}(\Sigma_\epsilon|\nu_\epsilon, \Sigma_{\epsilon,0}) \prod_{j=1}^{k} \text{InvW}(\Sigma_j|\nu_j, \Sigma_{j,0}), \end{aligned} \tag{3}$$

where $\Xi := (\Sigma^{-1} + \beta\Sigma_\epsilon^{-1})^{-1}$, and

$$\Sigma := \begin{pmatrix} \Sigma_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \Sigma_k \end{pmatrix}.$$

As before, the block structure of Σ is given by the clustering $\mathcal{C}$. The proposed model is the same model as in Eq. (1), with the main difference that the noise term $\beta\Sigma_\epsilon^{-1}$ is added to the precision matrix of the normal distribution.

$1 \gg \beta > 0$ is a hyper-parameter that is fixed to a small positive value accounting for the degree of noise on the precision matrix. Furthermore, we assume non-informative priors on Σ_j and Σ_ϵ by setting $\nu_j = d_j + 1, \Sigma_j = I_{d_j}$ and $\nu_\epsilon = d + 1, \Sigma_{\epsilon,0} = I_d$.

Remark on the parameterization We note that as an alternative parameterization, we could have defined $\Xi := (\Sigma^{-1} + \Sigma_\epsilon^{-1})^{-1}$, and instead place a prior on Σ_ϵ that encourages Σ_ϵ^{-1} to be small in terms of some matrix norm. For example, we could have set $\Sigma_{\epsilon,0} = \frac{1}{\beta} I_d$. We chose the parameterization $\Xi := (\Sigma^{-1} + \beta\Sigma_\epsilon^{-1})^{-1}$, since it allows us to set β to 0, which recovers the basic inverse Wishart prior model.

4.2 Estimation of the marginal likelihood

The marginal likelihood of the data given our proposed model can be expressed as follows:

$$\begin{aligned} &p(\mathbf{x}_1, \ldots, \mathbf{x}_n | \nu_\epsilon, \Sigma_{\epsilon,0}, \{\nu_j\}_j, \{\Sigma_{j,0}\}_j, \mathcal{C}) \\ &= \int \text{Normal}(\mathbf{x}_1, \ldots, \mathbf{x}_n|\mathbf{0}, \Xi) \\ &\cdot \prod_{j=1}^{k} \text{InvW}(\Sigma_j|\nu_j, \Sigma_{j,0}) d(\Sigma_j \succ 0) \\ &\cdot \text{InvW}(\Sigma_\epsilon|\nu_\epsilon, \Sigma_{\epsilon,0}) d(\Sigma_\epsilon \succ 0). \end{aligned}$$

where $\Xi := (\Sigma^{-1} + \beta\Sigma_\epsilon^{-1})^{-1}$.

Clearly, if $\beta = 0$, we recover the basic inverse Wishart prior model, as discussed in Sect. 3, and the marginal likelihood has a closed form solution due to the conjugacy of the covariance matrix of the Gaussian and the inverse Wishart prior. However, if $\beta > 0$, there is no analytic solution anymore. Therefore, we propose to either use an estimate based on a variational approximation (Sect. 4.2.2) or on MCMC (Sect. 4.2.3). Both of our estimates require the calculation of the maximum a posterior (MAP) solution which we explain first in Sect. 4.2.1.

Remark on BIC type approximation of the marginal likelihood We note that for our proposed model an approximation of the marginal likelihood using BIC is not sensible. To see this, recall that BIC consists of two terms: the data log-likelihood under the model with the maximum likelihood estimate, and a penalty depending on the number of free parameters. The maximum likelihood estimate is

$$\hat{\Sigma}, \hat{\Sigma}_\epsilon = \arg\max_{\Sigma, \Sigma_\epsilon} \sum_{i=1}^{n} \log \text{Normal}(\mathbf{x}_i|\mathbf{0}, (\Sigma^{-1} + \beta\Sigma_\epsilon^{-1})^{-1}),$$

where S is the sample covariance matrix. Note that without the specification of a prior, it is valid that $\hat{\Sigma}$, $\hat{\Sigma}_\epsilon$ are not positive definite as long as the matrix $\hat{\Sigma}^{-1} + \beta\hat{\Sigma}_\epsilon^{-1}$ is positive definite. Therefore, $\hat{\Sigma}^{-1} + \beta\hat{\Sigma}_\epsilon^{-1} = S^{-1}$, and the data likelihood under the model with the maximum likelihood estimate is simply $\sum_{i=1}^{n} \log \text{Normal}(\mathbf{x}_i|\mathbf{0}, S)$, which is independent of the clustering. Furthermore, the number of *free* parameters is $(d^2 - d)/2$ which is also independent of the clustering. That means, for any clustering we end up with the same BIC.

Furthermore, a Laplacian approximation as used in the generalized Bayesian information criterion (Konishi et al. 2004) is also not suitable, since in our case the parameter space is over the positive-definite matrices.

4.2.1 Calculation of maximum a posterior solution

Finding the exact MAP is crucial for the quality of the marginal likelihood approximation that we will describe later in Sects. 4.2.2 and 4.2.3. In this section, we explain in detail how the corresponding optimization problem can be solved with a 3-block ADMM method, which is guaranteed to converge to the global optimum.

First note that

$$\begin{aligned} &p(\Sigma, \Sigma_\epsilon|\mathbf{x}_1, \ldots, \mathbf{x}_n, \nu_\epsilon, \Sigma_{\epsilon,0}, \{\nu_j\}_j, \{\Sigma_{j,0}\}_j, \mathcal{C}) \\ &\quad \propto \text{Normal}(\mathbf{x}_1, \ldots, \mathbf{x}_n|\mathbf{0}, \Xi) \\ &\qquad \cdot \prod_{j=1}^{k} \text{InvW}(\Sigma_j|\nu_j, \Sigma_{j,0}) \\ &\qquad \cdot \text{InvW}(\Sigma_\epsilon|\nu_\epsilon, \Sigma_{\epsilon,0}) \end{aligned}$$

where $\Xi := (\Sigma^{-1} + \beta\Sigma_\epsilon^{-1})^{-1}$.

Therefore,

$$\begin{aligned} &\log p(\Sigma, \Sigma_\epsilon|\mathbf{x}_1, \ldots, \mathbf{x}_n, \nu_\epsilon, \Sigma_{\epsilon,0}, \{\nu_j\}_j, \{\Sigma_{j,0}\}_j, \mathcal{C}) \\ &\quad = -\frac{n}{2}\log|\Xi| - \frac{n}{2}\text{trace}(S\Xi^{-1}) \\ &\qquad - \frac{\nu_\epsilon + d + 1}{2}\log|\Sigma_\epsilon| - \frac{1}{2}\text{trace}(\Sigma_{\epsilon,0}\Sigma_\epsilon^{-1}) \\ &\qquad + \sum_{j=1}^{k}\left(-\frac{\nu_j + d_j + 1}{2}\log|\Sigma_j| - \frac{1}{2}\text{trace}(\Sigma_{j,0}\Sigma_j^{-1})\right) \\ &\qquad + const \\ &\quad = \frac{1}{2}\Big(n \cdot \log|\Xi^{-1}| - n \cdot \text{trace}(S\Xi^{-1}) \\ &\qquad + (\nu_\epsilon + d + 1) \cdot \log|\Sigma_\epsilon^{-1}| - \text{trace}(\Sigma_{\epsilon,0}\Sigma_\epsilon^{-1}) \\ &\qquad + \sum_{j=1}^{k}\Big((\nu_j + d_j + 1) \cdot \log|\Sigma_j^{-1}| - \text{trace}(\Sigma_{j,0}\Sigma_j^{-1})\Big)\Big) \\ &\qquad + const, \end{aligned}$$

where the constant is with respect to Σ_ϵ, $\Sigma_1, \ldots \Sigma_k$, and d_j denotes the number of variables in cluster j.

Solution using a 3-Block ADMM Finding the MAP can be formulated as a convex optimization problem by a change of parameterization: by defining $X := \Sigma^{-1}$, $X_j := \Sigma_j^{-1}$, and $X_\epsilon := \Sigma_\epsilon^{-1}$, we get the following convex optimization problem:

$$\begin{aligned} \underset{X \succ 0, X_\epsilon \succ 0}{\text{minimize}} \; & n \cdot \text{trace}(S(X + \beta X_\epsilon)) - n \cdot \log|X + \beta X_\epsilon| \\ & + \text{trace}(A_\epsilon X_\epsilon) - a_\epsilon \cdot \log|X_\epsilon| \\ & + \sum_{j=1}^{k}\Big(\text{trace}(A_j X_j) - a_j \cdot \log|X_j|\Big), \end{aligned} \tag{4}$$

where, for simplifying notation, we introduced the following constants:

$$\begin{aligned} A_\epsilon &:= \Sigma_{\epsilon,0}, \\ a_\epsilon &:= \nu_\epsilon + d + 1, \\ A_j &:= \Sigma_{j,0}, \\ a_j &:= \nu_j + d_j + 1. \end{aligned}$$

From this form, we see immediately that the problem is strictly convex jointly in X_ϵ and X.[1]

[1] Since $-log|X|$ is a strictly convex function and trace(XS) is a linear function.

We further reformulate the problem by introducing an additional variable Z:

$$\begin{aligned}&\text{minimize } f(X_\epsilon, X_1, \ldots, X_k, Z)\\ &\text{subject to}\\ &Z = X + \beta X_\epsilon,\\ &X_\epsilon, X_1, \ldots, X_k, Z \succeq 0,\end{aligned}$$

with

$$\begin{aligned}f(X_\epsilon, X_1, \ldots, X_k, Z) := n \cdot \text{trace}(SZ) - n \cdot \log|Z| \\ + \text{trace}(A_\epsilon X_\epsilon) - a_\epsilon \cdot \log|X_\epsilon| \\ + \sum_{j=1}^{k} \Big(\text{trace}(A_j X_j) - a_j \cdot \log|X_j|\Big).\end{aligned}$$

It is tempting to use a 2-Block ADMM algorithm, e.g., in Boyd et al. (2011), which leads to two optimization problems: update of X, X_ϵ and update of Z. However, unfortunately, in our case the resulting optimization problem for updating X, X_ϵ does not have an analytic solution. Therefore, instead, we suggest the use of a 3-Block ADMM, which updates the following sequence:

$$\begin{aligned}X^{t+1} &:= \underset{X_1,\ldots,X_k \succ 0}{\arg\min} \sum_{j=1}^{k} \Big(\text{trace}(A_j X_j) - a_j \cdot \log|X_j|\Big)\\ &\quad + \text{trace}(U^t(X + \beta X_\epsilon^t - Z^t))\\ &\quad + \frac{\rho}{2}||X + \beta X_\epsilon^t - Z^t||_F^2,\\ X_\epsilon^{t+1} &:= \underset{X_\epsilon \succ 0}{\arg\min}\ \text{trace}(A_\epsilon X_\epsilon) - a_\epsilon \cdot \log|X_\epsilon|\\ &\quad + \text{trace}(U^t(X^{t+1} + \beta X_\epsilon - Z^t))\\ &\quad + \frac{\rho}{2}||X^{t+1} + \beta X_\epsilon - Z^t||_F^2,\\ Z^{t+1} &:= \underset{Z \succ 0}{\arg\min}\ n \cdot \text{trace}(SZ) - n \cdot \log|Z|\\ &\quad + \text{trace}(U^t(X^{t+1} + \beta X_\epsilon^{t+1} - Z))\\ &\quad + \frac{\rho}{2}||X^{t+1} + \beta X_\epsilon^{t+1} - Z||_F^2,\\ U^{t+1} &:= \rho(X^{t+1} + \beta X_\epsilon^{t+1} - Z^{t+1}) + U^t,\end{aligned}$$

where U is the Lagrange multiplier, and X^t, Z^t, U^t, denotes X, Z, U at iteration t; $\rho > 0$ is the learning rate.[2]

Each of the above sub-optimization problem can be solved efficiently via the following strategy. The zero gradient condition for the first optimization problem with variable X is

$$-X_j^{-1} + \frac{\rho}{a_j} X_j = -\frac{1}{a_j}(A_j + U_j + \rho(\beta X_{\epsilon,j} - Z_j)).$$

The zero gradient condition for the second optimization problem with variable X_ϵ is

$$-X_\epsilon^{-1} + \frac{\rho\beta^2}{a_\epsilon} X_\epsilon = -\frac{1}{a_\epsilon}(A_\epsilon + \beta U + \rho\beta(X - Z)).$$

The zero gradient condition for the third optimization problem with variable Z is

$$-Z^{-1} + \frac{\rho}{n} Z = \frac{1}{n}(U - nS + \rho(X + \beta X_\epsilon)).$$

Each of the above three optimization problem can be solved via an eigenvalue decomposition as follows. We need to solve V such that it satisfies:

$$-V^{-1} + \lambda V = R \quad \wedge \quad V \succeq 0$$

Since R is a symmetric matrix (not necessarily positive or negative semi-definite), we have the eigenvalue decomposition:

$$QLQ^T = R,$$

where Q is an orthonormal matrix and L is a diagonal matrix with real values. Denoting $Y := Q^T V Q$, we have

$$-Y^{-1} + \lambda Y = L, \tag{5}$$

Since the solution Y must also be a diagonal matrix, we have $Y_{ij} = 0$, for $j \neq i$, and we must have that

$$-(Y_{ii})^{-1} + \lambda Y_{ii} = L_{ii}. \tag{6}$$

Then, Eq. (6) is equivalent to

$$\lambda Y_{ii}^2 - L_{ii} Y_{ii} - 1 = 0,$$

and therefore, one solution is

$$Y_{ii} = \frac{L_{ii} + \sqrt{L_{ii}^2 + 4\lambda}}{2\lambda}.$$

Note that for $\lambda > 0$, we have that $Y_{ii} > 0$. Therefore, we have that the resulting Y solves Eq. (5) and moreover

$$V = QYQ^T \succ 0.$$

That means, we can solve the semi-definite problem with only one eigenvalue decomposition, and therefore is in $O(d^3)$.

[2] In our experiments, we set the learning rate ρ initially to 1.0, and increase it every 100 iterations by a factor of 1.1. We found experimentally that this speeds-up the convergence of ADMM.

Finally, we note that in contrast to the 2-block ADMM, a general 3-block ADMM does not have a convergence guarantee for any $\rho > 0$. However, using a recent result from (Lin et al. 2018), we can show in "Appendix A" that in our case the conditions for convergence are met for any $\rho > 0$.

4.2.2 Variational approximation of the marginal likelihood

Here, we explain our strategy for the calculation of a variational approximation of the marginal likelihood. For simplicity, let $\boldsymbol{\theta}$ denote the vector of all parameters, $\mathcal{X}$ the observed data, and $\boldsymbol{\eta}$ the vector of all hyper-parameters.

Let $\hat{\boldsymbol{\theta}}$ denote the posterior mode. Furthermore, let $g(\boldsymbol{\theta})$ be an approximation of the posterior distribution $p(\boldsymbol{\theta}|\mathcal{X}, \boldsymbol{\eta}, \mathcal{C})$ that is accurate around the mode $\hat{\boldsymbol{\theta}}$.

Then, we have

$$\begin{aligned} p(\mathcal{X}|\boldsymbol{\eta}, \mathcal{C}) &= \frac{p(\boldsymbol{\theta}, \mathcal{X}|\boldsymbol{\eta}, \mathcal{C})}{p(\boldsymbol{\theta}|\mathcal{X}, \boldsymbol{\eta}, \mathcal{C})} \\ &= \frac{p(\hat{\boldsymbol{\theta}}, \mathcal{X}|\boldsymbol{\eta}, \mathcal{C})}{p(\hat{\boldsymbol{\theta}}|\mathcal{X}, \boldsymbol{\eta}, \mathcal{C})} \approx \frac{p(\hat{\boldsymbol{\theta}}, \mathcal{X}|\boldsymbol{\eta}, \mathcal{C})}{g(\hat{\boldsymbol{\theta}})}. \end{aligned} \tag{7}$$

Note that for the Laplace approximation we would use $g(\boldsymbol{\theta}) = N(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}, V)$, where V is an appropriate covariance matrix. However, here the posterior $p(\boldsymbol{\theta}|\mathcal{X}, \boldsymbol{\eta}, \mathcal{C})$ is a probability measure over the positive-definite matrices and not over $\mathbb{R}^d$, which makes the Laplace approximation inappropriate.

Instead, we suggest to approximate the posterior distribution $p(\Sigma_\epsilon, \Sigma_1, \ldots \Sigma_k|\mathbf{x}_1, \ldots, \mathbf{x}_n, \nu_\epsilon, \Sigma_{\epsilon,0}, \{\nu_j\}_j, \{\Sigma_{j,0}\}_j, \mathcal{C})$ by the factorized distribution

$$g := g_\epsilon(\Sigma_\epsilon) \cdot \prod_{j=1}^{k} g_j(\Sigma_j).$$

We define $g_\epsilon(\Sigma_\epsilon)$ and $g_j(\Sigma_j)$ as follows:

$$g_\epsilon(\Sigma_\epsilon) := \text{InvW}(\Sigma_\epsilon|\nu_{g,\epsilon}, \Sigma_{g,\epsilon}),$$

with

$$\Sigma_{g,\epsilon} := (\nu_{g,\epsilon} + d + 1) \cdot \hat{\Sigma}_\epsilon,$$

where $\hat{\Sigma}_\epsilon$ is the mode of the posterior probability $p(\Sigma_\epsilon|\mathcal{X}, \boldsymbol{\eta}, \mathcal{C})$ (as calculated in the previous section). Note that this choice ensures that the mode of g_ϵ is the same as the mode of $p(\Sigma_\epsilon|\mathbf{x}_1, \ldots, \mathbf{x}_n, \boldsymbol{\eta}, \mathcal{C})$. Analogously, we set

$$g_j(\Sigma_j) := \text{InvW}(\Sigma_j|\nu_{g,j}, \Sigma_{g,j}),$$

with

$$\Sigma_{g,j} := (\nu_{g,j} + d_j + 1) \cdot \hat{\Sigma}_j,$$

where $\hat{\Sigma}_j$ is the mode of the posterior probability $p(\Sigma_j|\mathcal{X}, \boldsymbol{\eta}, \mathcal{C})$. The remaining parameters $\nu_{g,\epsilon} \in \mathbb{R}$ and $\nu_{g,j} \in \mathbb{R}$ are optimized by minimizing the KL-divergence between the factorized distribution g and the posterior distribution $p(\Sigma_\epsilon, \Sigma_1, \ldots \Sigma_k|\mathbf{x}_1, \ldots, \mathbf{x}_n, \boldsymbol{\eta}, \mathcal{C})$. The details of the following derivations are given in "Appendix B". For simplicity, let us denote $g_J := \prod_{j=1}^{k} g_j$, then we have

$$\begin{aligned} KL(g||p) = &- \int g_\epsilon(\Sigma_\epsilon) \cdot \prod_{j=1}^{k} g_j(\Sigma_j) \\ &\log \frac{p(\Sigma_\epsilon, \Sigma_1, \ldots \Sigma_k, \mathbf{x}_1, \ldots, \mathbf{x}_n|\boldsymbol{\eta}, \mathcal{C})}{g_\epsilon(\Sigma_\epsilon) \cdot \prod_{j=1}^{k} g_j(\Sigma_j)} d\Sigma_\epsilon d\Sigma \\ &+ c \\ = &-\frac{1}{2} n \, \mathbb{E}_{g_J, g_\epsilon}[\log|\Sigma^{-1} + \beta\Sigma_\epsilon^{-1}|] \\ &+ \frac{1}{2}(\nu_\epsilon + d + 1) \, \mathbb{E}_{g_\epsilon}[\log|\Sigma_\epsilon|] \\ &+ \frac{1}{2}\text{trace}((\Sigma_{\epsilon,0} + \beta n S) \, \mathbb{E}_{g_\epsilon}[\Sigma_\epsilon^{-1}]) \\ &- \text{Entropy}[g_\epsilon] \\ &+ \frac{1}{2} \sum_{j=1}^{k} (\nu_j + d_j + 1) \, \mathbb{E}_{g_j}[\log|\Sigma_j|] \\ &+ \frac{1}{2} \sum_{j=1}^{k} \text{trace}((\Sigma_{j,0} + n S_j) \, \mathbb{E}_{g_j}[\Sigma_j^{-1}]) \\ &- \sum_{j=1}^{k} \text{Entropy}[g_j] + c, \end{aligned}$$

where c is a constant with respect to g_ϵ and g_j. However, the term $E_{g_J, g_\epsilon}[\log|\Sigma^{-1} + \beta\Sigma_\epsilon^{-1}|]$ cannot be solved analytically; therefore, we need to resort to some sort of approximation.

We assume that $E_{g_J, g_\epsilon}[\log|\Sigma^{-1} + \beta\Sigma_\epsilon^{-1}|] \approx E_{g_J, g_\epsilon}[\log|\Sigma^{-1}|]$. This way, we get

$$\begin{aligned} KL(g||p) \approx &\, KL(g_\epsilon \,||\, \text{InvW}(\nu_\epsilon, \Sigma_{\epsilon,0} + \beta n S)) \\ &+ \sum_{j=1}^{k} KL(g_j \,||\, \text{InvW}(\nu_j + n, \Sigma_{j,0} + n S_j)) \\ &+ c', \end{aligned}$$

where we used that

$$\mathbb{E}_{g_J,g_\epsilon}[\log|\Sigma^{-1}|] = -\sum_{j=1}^{k}\mathbb{E}_{g_j}[\log|\Sigma_j|],$$

and c' is a constant with respect to g_ϵ and g_j.

From the above expression, we see that we can optimize the parameters of g_ϵ and g_j independently from each other. The optimal parameter $\hat{\nu}_{g,\epsilon}$ for g_ϵ is

$$\begin{aligned}\hat{\nu}_{g,\epsilon} &= \arg\min_{\nu_{g,\epsilon}} KL(g_\epsilon \,||\, \mathrm{InvW}(\nu_\epsilon, \Sigma_{\epsilon,0}+\beta n S))\\ &= \arg\min_{\nu_{g,\epsilon}} \frac{\nu_{g,\epsilon}}{\nu_{g,\epsilon}+d+1}\mathrm{trace}\Big(\Big(\Sigma_{\epsilon,0}+\beta n S\Big)\hat{\Sigma}_\epsilon^{-1}\Big)\\ &\quad - 2\log\Gamma_d\left(\frac{\nu_{g,\epsilon}}{2}\right) - \nu_{g,\epsilon}d + d\nu_\epsilon\log(\nu_{g,\epsilon}+d+1)\\ &\quad + (\nu_{g,\epsilon}-\nu_\epsilon)\sum_{i=1}^{d}\psi\left(\frac{\nu_{g,\epsilon}-d+i}{2}\right).\end{aligned}$$

And analogously, we have

$$\begin{aligned}\hat{\nu}_{g,j} &= \arg\min_{\nu_{g,j}} \frac{\nu_{g,j}}{\nu_{g,j}+d_j+1}\mathrm{trace}\Big(\Big(\Sigma_{j,0}+nS_j\Big)\hat{\Sigma}_j^{-1}\Big)\\ &\quad - 2\log\Gamma_{d_j}\left(\frac{\nu_{g,j}}{2}\right) - \nu_{g,j}d_j\\ &\quad + d_j(\nu_j+n)\log(\nu_{g,j}+d_j+1)\\ &\quad + (\nu_{g,j}-\nu_j-n)\sum_{i=1}^{d_j}\psi\left(\frac{\nu_{g,j}-d_j+i}{2}\right).\end{aligned}$$

Each is a one-dimensional non-convex optimization problem that we solve with Brent's method (Brent 1971).

Discussion: Advantages over full variational approaches We described here an approximation to the marginal likelihood that can be considered as a blending of the ideas of the Laplace approximation (using the MAP) and a variational approximation where *all* parameters are learned by minimizing the Kullback–Leibler divergence between a variational distribution and the true posterior distribution. We refer to the latter as a *full* variational approximation. For simplicity, here, let us denote by Σ the positive-definite matrix for which we seek the posterior distribution, and let Σ_g denote the parameter matrix of the variational distribution.

An obvious limitation of the full variational approach is that the expectation involving Σ cannot be calculated analytically anymore. As a solution, recent works on black-box variational inference propose to use a Monte Carlo estimate of the expectation of the gradient. In order to address high variance of the estimator, several techniques have been proposed (e.g., control variates and Rao–Blackwellization) among which the reparameterization trick appears to be the most promising (Ranganath et al. 2014; Kingma and Welling 2013; Kucukelbir et al. 2017). In particular, Stan (Carpenter et al. 2017) provides a readily available implementation of the reparameterization trick (Kucukelbir et al. 2017) which is named automatic differentiation variational inference (ADVI). In ADVI, the transformation is $\Sigma_g := L^T L$ with L being a triangular matrix where each component is sampled from $N(0,1)$. And the matrix L is the parameter of the variational distribution that is optimized with stochastic gradient descent. However, note that this optimization problem is a stochastic non-convex problem. In contrast, finding the MAP is a non-stochastic convex optimization problem and the proposed solution has a guarantee of converging to the global minima. Apart from that, we note that a full variational approximation does not have any theoretic quality guarantees, including the case where $\beta \to 0$. In the general case, our approach also does not have such guarantees. However, in the special case where $\beta \to 0$, we know that the true posterior distribution is an inverse Wishart distribution and therefore matches our choice of the variational distribution.

4.2.3 MCMC estimation of marginal likelihood

As an alternative to the variational approximation, we investigate an MCMC estimation based on Chib's method (Chib 1995; Chib and Jeliazkov 2001).

To simplify the description, we introduce the following notations

$$\begin{aligned}&\boldsymbol{\theta}_1 := \Sigma_\epsilon,\\ &\boldsymbol{\theta}_2,\ldots,\boldsymbol{\theta}_{k+1} := \Sigma_1,\ldots,\Sigma_k.\end{aligned}$$

Furthermore, we define $\boldsymbol{\theta}_{<i} := \{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_{i-1}\}$ and $\boldsymbol{\theta}_{>i} := \{\boldsymbol{\theta}_{i+1},\ldots,\boldsymbol{\theta}_{k+1}\}$. For simplicity, we also suppress in the notation the explicit conditioning on the hyper-parameters $\boldsymbol{\eta}$ and the clustering $\mathcal{C}$, which are both fixed.

Following the strategy of Chib (1995), the marginal likelihood can be expressed as

$$\begin{aligned}p(\mathcal{X}) &= \frac{p(\hat{\boldsymbol{\theta}}_1,\ldots,\hat{\boldsymbol{\theta}}_{k+1},\mathcal{X})}{p(\hat{\boldsymbol{\theta}}_1,\ldots,\hat{\boldsymbol{\theta}}_{k+1}|\mathcal{X})}\\ &= \frac{p(\hat{\boldsymbol{\theta}}_1,\ldots,\hat{\boldsymbol{\theta}}_{k+1},\mathcal{X})}{\prod_{i=1}^{k+1}p(\hat{\boldsymbol{\theta}}_i|\mathcal{X},\hat{\boldsymbol{\theta}}_1\ldots,\hat{\boldsymbol{\theta}}_{i-1})}\end{aligned} \tag{8}$$

In order to approximate $p(\mathcal{X})$ with Eq. (8), we need to estimate $p(\hat{\boldsymbol{\theta}}_i|\mathcal{X},\hat{\boldsymbol{\theta}}_1,\ldots\hat{\boldsymbol{\theta}}_{i-1})$. First, note that we can express the value of the conditional posterior distribution at $\hat{\boldsymbol{\theta}}_i$, as follows (see Chib and Jeliazkov (2001), Section 2.3):

$$\begin{aligned}&p(\hat{\boldsymbol{\theta}}_i|\mathcal{X},\hat{\boldsymbol{\theta}}_1,\ldots\hat{\boldsymbol{\theta}}_{i-1})\\ &= \frac{\mathbb{E}_{\boldsymbol{\theta}_{\geq i}\sim p(\boldsymbol{\theta}_{\geq i}|\mathcal{X},\hat{\boldsymbol{\theta}}_{<i})}[\alpha(\boldsymbol{\theta}_i,\hat{\boldsymbol{\theta}}_i|\hat{\boldsymbol{\theta}}_{<i},\boldsymbol{\theta}_{>i})q_i(\hat{\boldsymbol{\theta}}_i)]}{\mathbb{E}_{\boldsymbol{\theta}_{\geq i}\sim p(\boldsymbol{\theta}_{>i}|\mathcal{X},\hat{\boldsymbol{\theta}}_{\leq i})q(\boldsymbol{\theta}_i)}[\alpha(\hat{\boldsymbol{\theta}}_i,\boldsymbol{\theta}_i|\hat{\boldsymbol{\theta}}_{<i},\boldsymbol{\theta}_{>i})]},\end{aligned} \tag{9}$$

where $q_i(\boldsymbol{\theta}_i)$ is a proposal distribution for $\boldsymbol{\theta}_i$, and the acceptance probability of moving from state $\boldsymbol{\theta}_i$ to state $\boldsymbol{\theta}_i'$, holding the other states fixed is defined as

$$\alpha(\boldsymbol{\theta}_i, \boldsymbol{\theta}_i' | \boldsymbol{\theta}_{<i}, \boldsymbol{\theta}_{>i}) := \min\left\{1, \frac{p(\mathscr{X}, \boldsymbol{\theta}_{<i}, \boldsymbol{\theta}_{>i}, \boldsymbol{\theta}_i') \cdot q_i(\boldsymbol{\theta}_i)}{p(\mathscr{X}, \boldsymbol{\theta}_{<i}, \boldsymbol{\theta}_{>i}, \boldsymbol{\theta}_i) \cdot q_i(\boldsymbol{\theta}_i')}\right\}. \tag{10}$$

Next, using Eq. (9), we can estimate $p(\hat{\boldsymbol{\theta}}_i | \mathscr{X}, \hat{\boldsymbol{\theta}}_1, \ldots \hat{\boldsymbol{\theta}}_{i-1})$ with a Monte Carlo approximation with M samples:

$$\begin{aligned} &p(\hat{\boldsymbol{\theta}}_i | \mathscr{X}, \hat{\boldsymbol{\theta}}_1, \ldots \hat{\boldsymbol{\theta}}_{i-1}) \\ &\approx \frac{\frac{1}{M}\sum_{m=1}^{M} \alpha\left(\boldsymbol{\theta}_i^{i,m}, \hat{\boldsymbol{\theta}}_i | \hat{\boldsymbol{\theta}}_{<i}, \boldsymbol{\theta}_{>i}^{i,m}\right) q_i(\hat{\boldsymbol{\theta}}_i)}{\frac{1}{M}\sum_{m=1}^{M} \alpha\left(\hat{\boldsymbol{\theta}}_i, \boldsymbol{\theta}_i^{q,m} | \hat{\boldsymbol{\theta}}_{<i}, \boldsymbol{\theta}_{>i}^{i+1,m}\right)} \end{aligned} \tag{11}$$

where $\boldsymbol{\theta}_i^{a,m} \sim p(\boldsymbol{\theta}_i | \mathscr{X}, \hat{\boldsymbol{\theta}}_{<a})$, $\boldsymbol{\theta}_{>i}^{a,m} \sim p(\boldsymbol{\theta}_{>i} | \mathscr{X}, \hat{\boldsymbol{\theta}}_{<a})$, and $\boldsymbol{\theta}_i^{q,m} \sim q(\boldsymbol{\theta}_i)$.

Finally, in order to sample from $p(\boldsymbol{\theta}_{\geq i} | \mathscr{X}, \hat{\boldsymbol{\theta}}_{<i})$, we propose to use the Metropolis–Hastings within Gibbs sampler as shown in Algorithm 1. $MH_j(\boldsymbol{\theta}_j^t, \boldsymbol{\psi})$ denotes the Metropolis–Hastings algorithm with current state $\boldsymbol{\theta}_j^t$, and acceptance probability $\alpha(\boldsymbol{\theta}_j, \boldsymbol{\theta}_j' | \boldsymbol{\psi})$, Eq. (10), and $\boldsymbol{\theta}_{\geq i}^0$ is a sample after the burn-in. For the proposal distribution $q_i(\boldsymbol{\theta}_i)$, we use

$$q_i := \begin{cases} \text{InvW}(\nu, \hat{\Sigma}_\epsilon \cdot (\nu + d + 1)) \\ \quad \text{with } \nu = \beta\kappa \cdot n + \nu_\epsilon & \text{if } i = 1, \\ \text{InvW}(\nu, \hat{\Sigma}_{i-1} \cdot (\nu + d_{i-1} + 1)) \\ \quad \text{with } \nu = (1-\beta)\kappa \cdot n + \nu_{i-1} & \text{else.} \end{cases} \tag{12}$$

Here, $\kappa > 0$ is a hyper-parameter of the MCMC algorithm that is chosen to control the acceptance probability. Note that if we choose $\kappa = 1$ and β is 0, then the proposal distribution $q_i(\boldsymbol{\theta}_i)$ equals the posterior distribution $p(\boldsymbol{\theta}_i | \mathscr{X}, \hat{\boldsymbol{\theta}}_1, \ldots \hat{\boldsymbol{\theta}}_{i-1})$. However, in practice, we found that the acceptance probabilities can be too small, leading to unstable estimates and division by 0 in Eq. (11). Therefore, for our experiments we chose $\kappa = 10$.

Algorithm 1 Metropolis–Hastings within Gibbs sampler for sampling from $p(\boldsymbol{\theta}_{\geq i} | \mathscr{X}, \hat{\boldsymbol{\theta}}_{<i})$.

for t from 1 to M **do**
  for j from i to $k+1$ **do**
    $\boldsymbol{\psi} := \{\hat{\boldsymbol{\theta}}_{<i}, \boldsymbol{\theta}_i^t, \ldots, \boldsymbol{\theta}_{j-1}^t, \boldsymbol{\theta}_{>j}^{t-1}\}$
    $\boldsymbol{\theta}_j^t := MH_j(\boldsymbol{\theta}_j^{t-1}, \boldsymbol{\psi})$
  end for
end for

4.3 Restricting the hypotheses space

The number of possible clusterings follows the Bell numbers, and therefore, it is infeasible to enumerate all possible clusterings, even if the number of variables d is small. It is therefore crucial to restrict the hypotheses space to a subset of all clusterings that are likely to contain the true clustering. We denote this subset as $\mathscr{C}^*$.

We suggest to use spectral clustering on different estimates of the precision matrix to acquire the set of clusterings $\mathscr{C}^*$. A motivation for this heuristic is given in "Appendix C".

First, for an appropriate λ, we estimate the precision matrix using

$$X^* := \underset{X \succeq 0}{\arg\min} - \log|X| + \text{trace}(XS) + \lambda \sum_{i \neq j} |X_{ij}|^q. \tag{13}$$

In our experiments, we take $q = 1$, which is equivalent to the Graphical Lasso (Friedman et al. 2008) with an ℓ_1-penalty on all entries of X except the diagonal. In the next step, we then construct the Laplacian L as defined in the following.

$$\begin{aligned} L_{ii} &= \sum_{k \neq i} |X_{ik}^*|^q, \\ L_{ij} &= -|X_{ij}^*|^q \quad \text{for } i \neq j. \end{aligned} \tag{14}$$

Finally, we use k-means clustering on the eigenvectors of the Laplacian L. The details of acquiring the set of clusterings $\mathscr{C}^*$ using the spectral clustering method are summarized below:

Algorithm 2 Spectral Clustering for variable clustering with the Gaussian graphical model.

J := set of regularization parameter values.
K_{max} := maximum number of considered clusters.
$\mathscr{C}^* := \{\}$
for $\lambda \in J$ **do**
  X^* := solve optimization problem from Eq. (13).
  $(\mathbf{e}_1, \ldots, \mathbf{e}_{K_{max}})$:= determine the eigenvectors corresponding to the K_{max} lowest eigenvalues of the Laplacian L as defined in Eq. (14).
  for $k \in \{2, \ldots, K_{max}\}$ **do**
    $\mathcal{C}_{\lambda,k}$:= cluster all variables into k partitions using k-means with $(\mathbf{e}_1, \ldots, \mathbf{e}_k)$.
    $\mathscr{C}^* := \mathscr{C}^* \cup \mathcal{C}_{\lambda,k}$
  end for
end for
return restricted hypotheses space $\mathscr{C}^*$

In Sect. 5.1 we confirm experimentally that, even in the presence of noise, $\mathscr{C}^*$ often contains the true clustering, or clusterings that are close to the true clustering.

Table 1 Evaluation of restricted hypotheses space for $d = 40$, $n \in \{20, 40, 400, 4000, 40{,}000, 4{,}000{,}000\}$

		20	40	400	4000	40,000	4,000,000
$\Sigma_j \sim \text{InvW}(d_j + 1, I_{d_j})$, no noise							
Spectral	ANMI	**0.77** (0.14)	**0.95** (0.06)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
	$\lvert\mathscr{C}^*\rvert$	140.8 (5.78)	139.0 (8.65)	112.8 (5.64)	99.8 (2.23)	101.4 (7.94)	98.4 (3.61)
Average	ANMI	0.38 (0.09)	0.38 (0.06)	0.45 (0.05)	0.45 (0.03)	0.45 (0.07)	0.45 (0.03)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
Single	ANMI	0.32 (0.08)	0.34 (0.09)	0.39 (0.08)	0.39 (0.08)	0.42 (0.14)	0.41 (0.08)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
$\Sigma_j \sim \text{InvW}(d_j + 1, I_{d_j})$, $\Sigma_\epsilon \sim \text{InvW}(d + 1, I_d)$, $\eta = 0.01$							
Spectral	ANMI	**0.49** (0.03)	**0.9** (0.03)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
	$\lvert\mathscr{C}^*\rvert$	143.2 (7.25)	144.4 (3.32)	108.6 (9.89)	105.4 (9.79)	103.6 (5.0)	97.0 (6.57)
Average	ANMI	0.26 (0.05)	0.34 (0.04)	0.46 (0.07)	0.51 (0.08)	0.42 (0.09)	0.45 (0.06)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
Single	ANMI	0.16 (0.08)	0.25 (0.08)	0.37 (0.03)	0.4 (0.06)	0.3 (0.12)	0.32 (0.09)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
$\Sigma_j \sim \text{InvW}(d_j + 1, I_{d_j})$, $\Sigma_\epsilon \sim \text{InvW}(d + 1, I_d)$, $\eta = 0.1$							
Spectral	ANMI	**0.34** (0.1)	**0.87** (0.09)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
	$\lvert\mathscr{C}^*\rvert$	121.4 (7.34)	106.4 (18.51)	35.4 (5.12)	33.2 (11.48)	37.4 (5.54)	31.0 (8.65)
Average	ANMI	0.1 (0.05)	0.15 (0.03)	0.34 (0.08)	0.37 (0.1)	0.26 (0.11)	0.28 (0.09)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
Single	ANMI	0.04 (0.03)	0.08 (0.04)	0.19 (0.11)	0.21 (0.06)	0.11 (0.03)	0.13 (0.02)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
$\Sigma_j \sim \text{Uniform}_{d_j}$, no noise							
Spectral	ANMI	**0.34** (0.1)	**0.87** (0.09)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
	$\lvert\mathscr{C}^*\rvert$	121.4 (7.34)	106.4 (18.51)	35.4 (5.12)	33.2 (11.48)	37.4 (5.54)	31.0 (8.65)
Average	ANMI	0.1 (0.06)	0.26 (0.07)	0.92 (0.11)	**1.0** (0.0)	**1.0** (0.0)	0.99 (0.03)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
Single	ANMI	0.04 (0.02)	0.13 (0.08)	0.82 (0.25)	**1.0** (0.0)	**1.0** (0.0)	0.99 (0.03)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
$\Sigma_j \sim \text{Uniform}_{d_j}$, $\Sigma_\epsilon \sim \text{Uniform}_d$, $\eta = 0.01$							
Spectral	ANMI	**0.28** (0.06)	**0.81** (0.1)	**0.94** (0.06)	**0.99** (0.03)	**0.99** (0.03)	**0.97** (0.03)
	$\lvert\mathscr{C}^*\rvert$	127.2 (3.6)	106.0 (5.29)	48.2 (9.77)	50.2 (5.95)	51.0 (8.94)	48.0 (5.69)
Average	ANMI	0.14 (0.05)	0.22 (0.04)	0.81 (0.16)	0.89 (0.1)	0.87 (0.12)	0.94 (0.12)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
Single	ANMI	0.04 (0.02)	0.1 (0.04)	0.78 (0.13)	0.71 (0.23)	0.78 (0.11)	0.79 (0.17)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
$\Sigma_j \sim \text{Uniform}_{d_j}$, $\Sigma_\epsilon \sim \text{Uniform}_d$, $\eta = 0.1$							
Spectral	ANMI	**0.3** (0.03)	**0.72** (0.08)	**0.88** (0.07)	**0.9** (0.07)	**0.87** (0.11)	0.88 (0.04)
	$\lvert\mathscr{C}^*\rvert$	126.2 (2.23)	120.4 (9.35)	74.4 (19.41)	87.2 (7.93)	79.2 (13.61)	77.0 (14.25)
Average	ANMI	0.08 (0.04)	0.26 (0.11)	0.83 (0.15)	0.88 (0.12)	**0.87** (0.11)	**0.94** (0.12)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
Single	ANMI	0.05 (0.03)	0.13 (0.07)	0.7 (0.14)	0.69 (0.15)	0.76 (0.12)	0.76 (0.14)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)

Ground truth contains 4 balanced clusters. Shows the oracle performance measured by ANMI for spectral clustering, average linkage, and single linkage. Note that that an ANMI score of 1.0 means that the true clustering is contained in the hypotheses space found by the clustering method. The size of the hypotheses space restricted by each clustering method is denoted by $\lvert\mathscr{C}^*\rvert$. Average results over 5 runs with standard deviation in brackets

The best ANMI scores are highlighted in bold

Table 2 Same setting as in Table 1 but with unbalanced clusters

		20	40	400	4000	40,000	4,000,000
$\Sigma_j \sim \text{InvW}(d_j + 1, I_{d_j})$, no noise							
Spectral	ANMI	**0.52** (0.13)	**0.85** (0.11)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
	$\lvert\mathscr{C}^*\rvert$	141.2 (6.62)	133.2 (8.03)	80.8 (8.21)	73.4 (8.89)	62.0 (7.38)	62.6 (7.23)
Average	ANMI	0.34 (0.06)	0.39 (0.05)	0.37 (0.04)	0.38 (0.07)	0.38 (0.06)	0.44 (0.09)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
Single	ANMI	0.33 (0.05)	0.35 (0.03)	0.32 (0.04)	0.32 (0.14)	0.27 (0.13)	0.39 (0.12)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
$\Sigma_j \sim \text{InvW}(d_j + 1, I_{d_j})$, $\Sigma_\epsilon \sim \text{InvW}(d + 1, I_d)$, $\eta = 0.01$							
Spectral	ANMI	**0.55** (0.13)	**0.81** (0.07)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
	$\lvert\mathscr{C}^*\rvert$	148.8 (4.62)	136.0 (6.81)	80.4 (9.77)	68.8 (10.3)	67.0 (5.93)	63.0 (14.3)
Average	ANMI	0.34 (0.06)	0.37 (0.08)	0.53 (0.12)	0.5 (0.1)	0.46 (0.1)	0.52 (0.1)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
Single	ANMI	0.29 (0.07)	0.29 (0.08)	0.41 (0.17)	0.4 (0.14)	0.37 (0.11)	0.32 (0.12)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
$\Sigma_j \sim \text{InvW}(d_j + 1, I_{d_j})$, $\Sigma_\epsilon \sim \text{InvW}(d + 1, I_d)$, $\eta = 0.1$							
Spectral	ANMI	**0.26** (0.04)	**0.5** (0.06)	**0.93** (0.07)	**0.93** (0.07)	**0.99** (0.02)	**0.91** (0.08)
	$\lvert\mathscr{C}^*\rvert$	144.4 (5.54)	159.2 (1.83)	121.0 (10.43)	120.2 (6.62)	117.0 (3.41)	113.2 (11.91)
Average	ANMI	0.2 (0.03)	0.22 (0.06)	0.37 (0.09)	0.36 (0.08)	0.41 (0.13)	0.44 (0.07)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
Single	ANMI	0.2 (0.08)	0.2 (0.07)	0.24 (0.04)	0.29 (0.05)	0.33 (0.07)	0.32 (0.05)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
$\Sigma_j \sim \text{Uniform}_{d_j}$, no noise							
Spectral	ANMI	**0.36** (0.06)	**0.72** (0.13)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
	$\lvert\mathscr{C}^*\rvert$	124.0 (7.29)	115.8 (9.89)	40.8 (12.5)	39.4 (5.2)	33.2 (4.79)	38.6 (5.24)
Average	ANMI	0.09 (0.04)	0.05 (0.08)	0.12 (0.07)	0.29 (0.07)	0.37 (0.07)	0.34 (0.14)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
Single	ANMI	0.01 (0.04)	0.0 (0.0)	0.0 (0.01)	0.06 (0.1)	0.17 (0.19)	0.13 (0.12)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
$\Sigma_j \sim \text{Uniform}_{d_j}$, $\Sigma_\epsilon \sim \text{Uniform}_d$, $\eta = 0.01$							
Spectral	ANMI	**0.39** (0.04)	**0.67** (0.11)	**0.85** (0.05)	**0.89** (0.07)	**0.87** (0.07)	**0.89** (0.06)
	$\lvert\mathscr{C}^*\rvert$	125.6 (8.06)	115.0 (12.85)	42.6 (7.09)	59.2 (11.55)	53.2 (9.2)	54.0 (6.69)
Average	ANMI	0.04 (0.03)	0.06 (0.05)	0.12 (0.06)	0.21 (0.08)	0.18 (0.09)	0.21 (0.13)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
Single	ANMI	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.02)	0.01 (0.05)	0.02 (0.05)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
$\Sigma_j \sim \text{Uniform}_{d_j}$, $\Sigma_\epsilon \sim \text{Uniform}_d$, $\eta = 0.1$							
Spectral	ANMI	**0.32** (0.06)	**0.68** (0.13)	**0.8** (0.09)	**0.81** (0.09)	**0.79** (0.07)	**0.78** (0.09)
	$\lvert\mathscr{C}^*\rvert$	124.2 (9.33)	109.6 (12.63)	66.6 (10.71)	74.2 (7.14)	62.8 (5.11)	65.2 (13.85)
Average	ANMI	0.04 (0.03)	0.06 (0.05)	0.09 (0.05)	0.19 (0.05)	0.13 (0.06)	0.2 (0.13)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
Single	ANMI	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.02)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)

Ground truth is 4 clusters with sizes 20, 10, 5, 5
The best ANMI scores are highlighted in bold

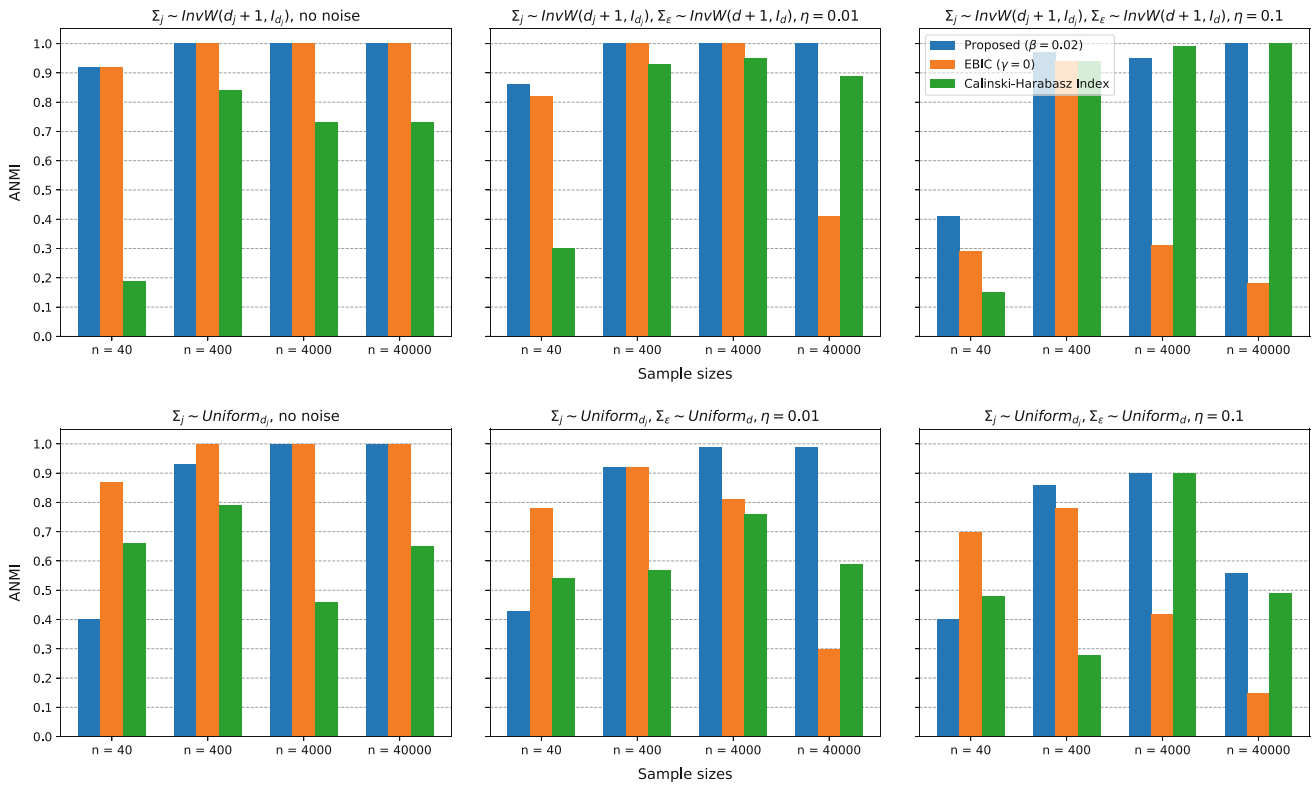

Fig. 1 The ANMI scores of the clustering selected by the proposed method (blue), EBIC (orange), and Calinski–Harabasz Index (green) on synthetic data sets with $d = 40$ and ground truth being 4 *balanced* clusters. Upper row and lower row shows results where the true precision matrix was generated from an inverse Wishart distribution, and a uniform distribution, respectively. No noise setting (left column), small noise (middle column), large noise (right column). ANMI score of 0.0 means correspondence with true clustering at pure chance level and 1.0 means perfect correspondence. In both settings, with and without noise, the proposed method tends to be among the best. In contrast, EBIC tends to suffer in the noise setting for large n and Calinski–Harabasz Index performs sub-optimal in the no noise setting. (Color figure online)

4.3.1 Posterior distribution over number of clusters

In principle, the posterior distribution for the number of clusters can be calculated using

$$p(k|\mathcal{X}) \propto \sum_{\mathcal{C}\in\mathscr{C}_k} p(\mathcal{X}|\mathcal{C}),$$

where $\mathscr{C}_k$ denotes the set of all clusterings with number of clusters being equal to k. Since this is computationally infeasible, we use the following approximation

$$P(k|X) \propto \sum_{\mathcal{C}\in\mathscr{C}_k} p(X|\mathcal{C}) \approx \sum_{\mathcal{C}\in\mathscr{C}_k^*} p(X|\mathcal{C}),$$

where $\mathscr{C}_k^*$ is the set of all clusterings with k clusters that are in the restricted hypotheses space $\mathscr{C}^*$.

5 Simulation study

In this section, we evaluate the proposed method on simulated data for which the ground truth is available. In Sect. 5.1, we evaluate the quality of the restricted hypotheses space $\mathscr{C}^*$, followed by Sect. 5.2, where we evaluated the proposed method's ability to select the best clustering in $\mathscr{C}^*$.

For the number of clusters, we consider the range from 2 to 15. For the set of regularization parameters of the spectral clustering method, we use $J := \{0.0001, 0.0005, 0.001, 0.002, 0.003, 0.004, 0.005, 0.006, 0.007, 0.008, 0.009, 0.01\}$ (see Algorithm 2).

In all experiments, the number of variables is $d = 40$, and the ground truth is 4 clusters with 10 variables each.

For generating positive-definite covariance matrices, we consider the following two distributions: $\text{InvW}(d+1, I_d)$, and Uniform_d, with dimension d. We denote by $U \sim \text{Uniform}_d$ the positive-definite matrix generated in the following way

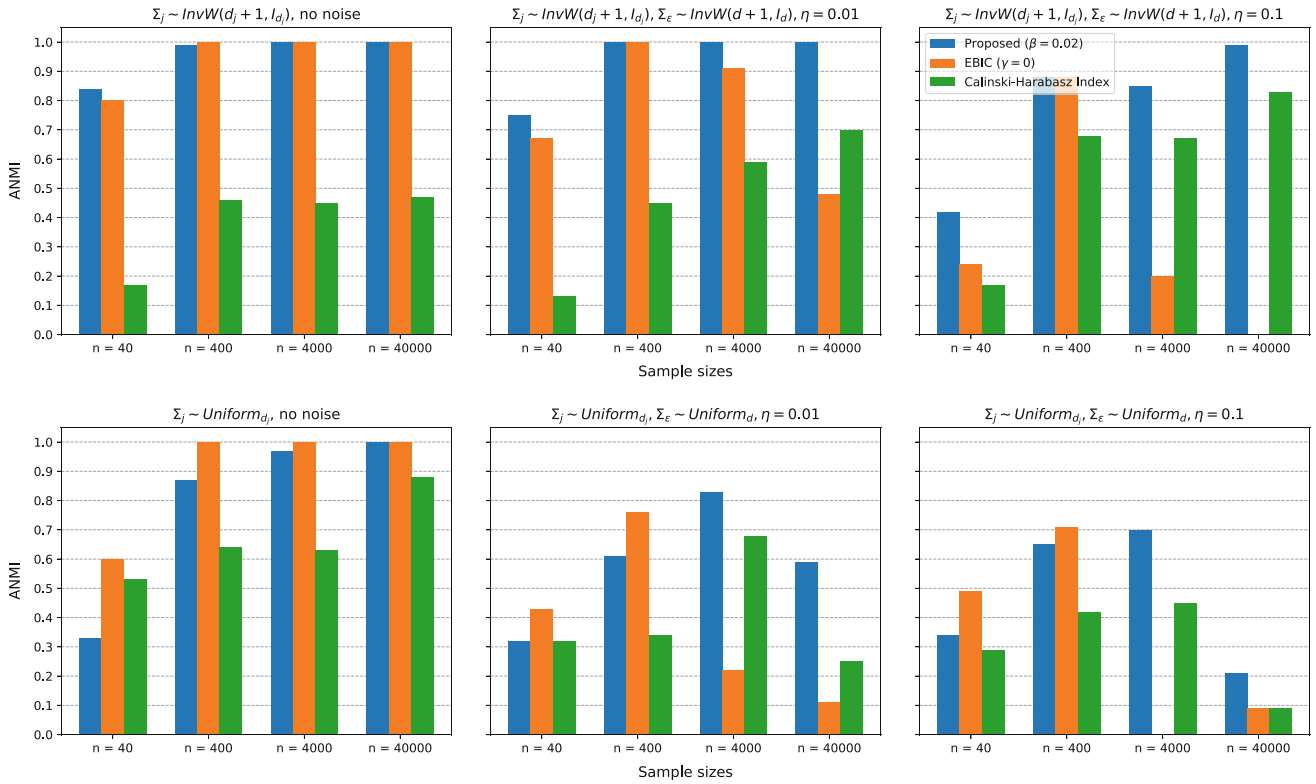

Fig. 2 Same settings as in Fig. 1, but ground truth being 4 *unbalanced* clusters

$$U = A + (0.001 - \lambda_{min}(A))I_d,$$

where $\lambda_{min}(A)$ is the smallest eigenvalue of A, and A is drawn as follows:

$$A_{i,j} = A_{j,i} \sim \text{Uniform}(-1, 1)\,, i \neq j$$
$$A_{i,i} = 0.$$

For generating Σ, we either sample each block j from $\text{InvW}(d_j + 1, I_{d_j})$ or from Uniform_{d_j}.

For generating the noise matrix Σ_ϵ, we sample either from $\text{InvW}(d + 1, I_d)$ or from Uniform_d. The final data are then sampled as follows:

$$x \sim N(0, (\Sigma^{-1} + \eta\Sigma_\epsilon^{-1})^{-1}),$$

where η defines the noise level.

For evaluation we use the adjusted normalized mutual information (ANMI), where 0.0 means that any correspondence with the true labels is at chance level, and 1.0 means that a perfect one-to-one correspondence exists (Vinh et al. 2010). We repeated all experiments 5 times and report the average ANMI score.

5.1 Evaluation of the restricted hypotheses space

First, independent of any model selection criteria, we check here the quality of the clusterings that are found with the spectral clustering algorithm from Sect. 4.3. We also compare to single and average linkage clustering as used in (Tan et al. 2015).

The set of all clusterings that are found is denoted by $\mathscr{C}^*$ (the restricted hypotheses space).

In order to evaluate the quality of the restricted hypotheses space $\mathscr{C}^*$, we report the oracle performance calculated by $\max_{\mathcal{C}\in\mathscr{C}^*} \text{ANMI}(\mathcal{C}, \mathcal{C}_T)$, where $\mathcal{C}_T$ denotes the true clustering, and $\text{ANMI}(\mathcal{C}, \mathcal{C}_T)$ denotes the ANMI score when comparing clustering $\mathcal{C}$ with the true clustering. In particular, a score of 1.0 means that the true clustering is contained in $\mathscr{C}^*$.

The results of all experiments with noise level $\eta \in \{0.0, 0.01, 0.1\}$ are shown in Table 1, for balanced clusters, and Table 2, for unbalanced clusters.

From these results, we see that the restricted hypotheses space of spectral clustering is around 100, considerably smaller than the number of all possible clusterings. More importantly, we also see that that $\mathscr{C}^*$ acquired by spectral clustering either contains the true clustering or a clustering that is close to the truth. In contrast, the hypotheses space

Table 3 Evaluation of clustering results for $d = 40$, $n \in \{20, 40, 400, 4000, 40{,}000, 4{,}000{,}000\}$

	20	40	400	4000	40,000	4,000,000
$\Sigma_j \sim \text{InvW}(d_j + 1, I_{d_j})$, no noise						
Proposed ($\beta = 0.01$)	**0.76** (0.14)	**0.93** (0.09)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
Proposed ($\beta = 0.02$)	0.7 (0.2)	0.92 (0.08)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
Proposed ($\beta = 0.03$)	0.67 (0.18)	0.88 (0.14)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
Basic inverse Wishart prior	0.73 (0.17)	**0.93** (0.09)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
EBIC ($\gamma = 0$)	0.12 (0.15)	0.92 (0.08)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
EBIC ($\gamma = 0.5$)	0.36 (0.03)	0.51 (0.04)	0.99 (0.03)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
EBIC ($\gamma = 1.0$)	0.35 (0.02)	0.39 (0.05)	0.96 (0.05)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
AIC	0.12 (0.15)	0.6 (0.49)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
Calinski–Harabasz Index	0.32 (0.03)	0.19 (0.16)	0.84 (0.13)	0.73 (0.0)	0.73 (0.0)	0.73 (0.0)
CGL (ALC)	0.06 (0.05)	0.03 (0.05)	0.11 (0.06)	0.04 (0.04)	0.06 (0.03)	0.06 (0.07)
DPVC	0.53 (0.07)	0.61 (0.17)	0.82 (0.06)	0.93 (0.09)	NA	NA
$\Sigma_j \sim \text{Uniform}_{d_j}$, no noise						
Proposed ($\beta = 0.01$)	0.12 (0.04)	0.48 (0.07)	0.94 (0.06)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
Proposed ($\beta = 0.02$)	0.12 (0.05)	0.4 (0.04)	0.93 (0.06)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
Proposed ($\beta = 0.03$)	0.12 (0.05)	0.39 (0.03)	0.93 (0.06)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
Basic inverse Wishart prior	0.14 (0.05)	0.76 (0.1)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
EBIC ($\gamma = 0$)	0.07 (0.04)	**0.87** (0.09)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
EBIC ($\gamma = 0.5$)	0.11 (0.05)	0.48 (0.12)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
EBIC ($\gamma = 1.0$)	0.11 (0.05)	0.38 (0.05)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
AIC	0.07 (0.04)	0.66 (0.34)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
Calinski–Harabasz Index	**0.15** (0.05)	0.66 (0.16)	0.79 (0.11)	0.46 (0.14)	0.65 (0.23)	0.59 (0.17)
CGL (ALC)	0.03 (0.02)	0.02 (0.02)	0.37 (0.03)	0.39 (0.0)	0.39 (0.0)	0.51 (0.25)
DPVC	0.01 (0.02)	0.03 (0.03)	0.4 (0.2)	0.51 (0.22)	NA	NA

Ground truth is 4 balanced clusters. Shows the ANMI of the selected models (standard deviation in brackets). No noise is added
The best ANMI scores are highlighted in bold

restricted by single and average linkage is smaller, but more often misses the true clustering.

5.2 Evaluation of clustering selection criteria

Here, we evaluate the performance of our proposed method for selecting the correct clustering in the restricted hypotheses space $\mathscr{C}^*$. We compare our proposed method (variational) with several baselines and two previously proposed methods (Tan et al. 2015; Palla et al. 2012). Except for the two previously proposed methods, we created $\mathscr{C}^*$ with the spectral clustering algorithm from Sect. 4.3.

As a cluster selection criteria, we compare our method to the extended Bayesian information criterion (EBIC) with $\gamma \in \{0, 0.5, 1\}$ (Chen and Chen 2008; Foygel and Drton 2010), Akaike information criteria (Akaike 1973), and the Calinski–Harabasz Index (CHI) (Caliński and Harabasz 1974). Note that EBIC and AIC are calculated based on the basic Gaussian graphical model (i.e., the model in Eq. 1, but ignoring the prior specification).[3] Furthermore, we note that EBIC is model consistent, and therefore, assuming that the true precision matrix contains nonzero entries in each element, will choose asymptotically the clustering that has only one cluster with all variables in it. However, as an advantage for EBIC, we exclude that clustering. Furthermore, we note that in contrast to EBIC and AIC, the Calinski–Harabasz Index is not a model-based cluster evaluation criterion. The Calinski–Harabasz Index is an heuristic that uses as clustering criterion the ratio of the variance within and across clusters. As such it is expected to give reasonable clustering results if the noise is considerably smaller in magnitude than the within-cluster variable partial correlations.

We remark that EBIC and AIC is not well defined if the sample covariance matrix is singular, in particular if $n < d$ or $n \approx d$. As an ad hoc remedy, which works well in

[3] As discussed in Sect. 4.2, EBIC (and also AIC) cannot be used with our proposed model.

Table 4 Evaluation of clustering results with $d = 40$, $n \in \{20, 40, 400, 4000, 40{,}000, 4{,}000{,}000\}$

	20	40	400	4000	40,000	4,000,000
$\Sigma_j \sim \text{InvW}(d_j + 1, I_{d_j})$, $\Sigma_\epsilon \sim \text{InvW}(d + 1, I_d)$, $\eta = 0.01$						
Proposed ($\beta = 0.01$)	0.44 (0.07)	0.86 (0.06)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
Proposed ($\beta = 0.02$)	0.41 (0.06)	0.86 (0.06)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	0.99 (0.03)
Proposed ($\beta = 0.03$)	0.38 (0.06)	0.8 (0.06)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	0.99 (0.03)
Basic inverse Wishart prior	**0.45** (0.07)	**0.89** (0.02)	**1.0** (0.0)	**1.0** (0.0)	0.41 (0.04)	0.39 (0.0)
EBIC ($\gamma = 0$)	0.02 (0.02)	0.82 (0.07)	**1.0** (0.0)	**1.0** (0.0)	0.41 (0.04)	0.39 (0.0)
EBIC ($\gamma = 0.5$)	0.25 (0.08)	0.32 (0.07)	0.98 (0.04)	**1.0** (0.0)	0.48 (0.13)	0.39 (0.0)
EBIC ($\gamma = 1.0$)	0.23 (0.07)	0.32 (0.07)	0.96 (0.06)	**1.0** (0.0)	0.66 (0.14)	0.39 (0.0)
AIC	0.0 (0.01)	0.54 (0.44)	**1.0** (0.0)	0.39 (0.0)	0.41 (0.04)	0.39 (0.0)
Calinski–Harabasz Index	0.26 (0.09)	0.3 (0.16)	0.93 (0.1)	0.95 (0.11)	0.89 (0.13)	0.84 (0.13)
CGL (ALC)	0.01 (0.02)	0.02 (0.05)	0.04 (0.05)	0.03 (0.02)	0.05 (0.06)	0.02 (0.02)
DPVC	0.33 (0.07)	0.42 (0.08)	0.59 (0.16)	0.21 (0.18)	NA	NA
$\Sigma_j \sim \text{InvW}(d_j + 1, I_{d_j})$, $\Sigma_\epsilon \sim \text{InvW}(d + 1, I_d)$, $\eta = 0.1$						
Proposed ($\beta = 0.01$)	0.1 (0.1)	0.4 (0.09)	0.93 (0.1)	0.39 (0.0)	0.33 (0.17)	0.29 (0.15)
Proposed ($\beta = 0.02$)	0.13 (0.09)	**0.41** (0.07)	**0.97** (0.04)	0.95 (0.11)	**1.0** (0.0)	**0.99** (0.03)
Proposed ($\beta = 0.03$)	0.13 (0.09)	0.4 (0.09)	0.95 (0.04)	**0.99** (0.03)	**1.0** (0.0)	**0.99** (0.03)
Basic inverse Wishart prior	0.1 (0.1)	0.4 (0.09)	0.93 (0.1)	0.23 (0.19)	0.18 (0.21)	0.23 (0.19)
EBIC ($\gamma = 0$)	0.09 (0.09)	0.29 (0.06)	0.94 (0.05)	0.31 (0.15)	0.18 (0.21)	0.23 (0.19)
EBIC ($\gamma = 0.5$)	0.12 (0.05)	0.2 (0.02)	0.87 (0.02)	0.41 (0.04)	0.18 (0.21)	0.23 (0.19)
EBIC ($\gamma = 1.0$)	**0.14** (0.06)	0.2 (0.02)	0.54 (0.07)	0.86 (0.24)	0.18 (0.21)	0.23 (0.19)
AIC	0.0 (0.0)	0.0 (0.01)	0.09 (0.15)	0.23 (0.19)	0.18 (0.21)	0.23 (0.19)
Calinski–Harabasz Index	0.11 (0.05)	0.15 (0.13)	0.94 (0.05)	**0.99** (0.03)	**1.0** (0.0)	**0.99** (0.03)
CGL (ALC)	0.02 (0.03)	0.0 (0.01)	0.01 (0.01)	0.01 (0.02)	0.0 (0.0)	0.0 (0.0)
DPVC	0.11 (0.06)	0.16 (0.06)	0.27 (0.06)	0.04 (0.04)	NA	NA
$\Sigma_j \sim \text{Uniform}_{d_j}$, $\Sigma_\epsilon \sim \text{Uniform}_d$, $\eta = 0.01$						
Proposed ($\beta = 0.01$)	0.1 (0.04)	0.45 (0.05)	0.92 (0.06)	**0.99** (0.03)	**0.99** (0.03)	**0.93** (0.1)
Proposed ($\beta = 0.02$)	0.12 (0.03)	0.43 (0.06)	0.92 (0.06)	**0.99** (0.03)	**0.99** (0.03)	**0.93** (0.1)
Proposed ($\beta = 0.03$)	0.13 (0.02)	0.39 (0.03)	0.89 (0.07)	**0.99** (0.03)	**0.99** (0.03)	**0.93** (0.1)
Basic inverse Wishart prior	0.11 (0.06)	0.65 (0.12)	**0.94** (0.06)	0.88 (0.12)	0.3 (0.28)	0.46 (0.14)
EBIC ($\gamma = 0$)	0.06 (0.04)	**0.78** (0.14)	0.92 (0.1)	0.81 (0.23)	0.3 (0.28)	0.46 (0.14)
EBIC ($\gamma = 0.5$)	0.1 (0.03)	0.44 (0.06)	**0.94** (0.06)	**0.99** (0.03)	0.3 (0.28)	0.46 (0.14)
EBIC ($\gamma = 1.0$)	0.1 (0.03)	0.39 (0.03)	**0.94** (0.06)	**0.99** (0.03)	0.3 (0.28)	0.46 (0.14)
AIC	0.06 (0.04)	0.24 (0.33)	0.35 (0.43)	0.44 (0.15)	0.3 (0.28)	0.46 (0.14)
Calinski–Harabasz Index	**0.14** (0.06)	0.54 (0.33)	0.57 (0.35)	0.76 (0.21)	0.59 (0.29)	0.66 (0.14)
CGL (ALC)	0.0 (0.01)	0.01 (0.01)	0.24 (0.18)	0.39 (0.0)	0.35 (0.08)	0.39 (0.0)
DPVC	0.0 (0.01)	0.06 (0.07)	0.29 (0.22)	0.44 (0.2)	NA	NA
$\Sigma_j \sim \text{Uniform}_{d_j}$, $\Sigma_\epsilon \sim \text{Uniform}_d$, $\eta = 0.1$						
Proposed ($\beta = 0.01$)	0.11 (0.02)	0.45 (0.05)	**0.88** (0.07)	0.79 (0.21)	**0.56** (0.34)	**0.64** (0.22)
Proposed ($\beta = 0.02$)	0.14 (0.04)	0.4 (0.02)	0.86 (0.07)	**0.9** (0.07)	**0.56** (0.34)	**0.64** (0.22)
Proposed ($\beta = 0.03$)	0.14 (0.04)	0.39 (0.03)	0.86 (0.07)	**0.9** (0.07)	**0.56** (0.34)	**0.64** (0.22)
Basic inverse Wishart prior	0.13 (0.04)	0.52 (0.07)	**0.88** (0.07)	0.42 (0.33)	0.15 (0.19)	0.23 (0.19)
EBIC ($\gamma = 0$)	0.12 (0.06)	**0.7** (0.1)	0.78 (0.22)	0.42 (0.33)	0.15 (0.19)	0.16 (0.19)
EBIC ($\gamma = 0.5$)	0.13 (0.04)	0.44 (0.05)	**0.88** (0.07)	0.48 (0.26)	0.15 (0.19)	0.16 (0.19)

Table 4 continued

	20	40	400	4000	40,000	4,000,000
EBIC ($\gamma = 1.0$)	0.12 (0.05)	0.39 (0.03)	**0.88** (0.07)	0.6 (0.3)	0.15 (0.19)	0.16 (0.19)
AIC	0.12 (0.06)	0.2 (0.17)	0.06 (0.12)	0.42 (0.33)	0.15 (0.19)	0.16 (0.19)
Calinski–Harabasz Index	**0.17** (0.06)	0.48 (0.29)	0.28 (0.34)	**0.9** (0.07)	0.49 (0.27)	0.63 (0.22)
CGL (ALC)	0.01 (0.01)	0.07 (0.08)	0.31 (0.15)	0.39 (0.0)	0.33 (0.11)	0.38 (0.02)
DPVC	0.0 (0.0)	0.1 (0.09)	0.35 (0.12)	0.19 (0.18)	NA	NA

Ground truth is 4 balanced clusters. Shows the ANMI of the selected models (standard deviation in brackets). Noise is added to the precision matrix
The best ANMI scores are highlighted in bold

Table 5 Evaluation of clustering results for $d = 40$, $n \in \{20, 40, 400, 4000, 40{,}000, 4{,}000{,}000\}$

	20	40	400	4000	40,000	4,000,000
$\Sigma_j \sim \text{InvW}(d_j + 1, I_{d_j})$, no noise						
Proposed ($\beta = 0.01$)	0.49 (0.15)	**0.84** (0.11)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
Proposed ($\beta = 0.02$)	0.47 (0.17)	**0.84** (0.11)	0.99 (0.02)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
Proposed ($\beta = 0.03$)	0.42 (0.19)	0.82 (0.13)	0.99 (0.02)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
Basic inverse Wishart prior	0.5 (0.15)	**0.84** (0.12)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
EBIC ($\gamma = 0$)	0.2 (0.17)	0.8 (0.12)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
EBIC ($\gamma = 0.5$)	0.24 (0.05)	0.37 (0.05)	0.99 (0.02)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
EBIC ($\gamma = 1.0$)	0.23 (0.06)	0.32 (0.04)	0.99 (0.02)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
AIC	0.15 (0.19)	0.16 (0.12)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
Calinski–Harabasz Index	0.17 (0.09)	0.17 (0.23)	0.46 (0.27)	0.45 (0.23)	0.47 (0.19)	0.4 (0.14)
CGL (ALC)	0.07 (0.11)	0.03 (0.04)	0.05 (0.07)	0.03 (0.03)	0.07 (0.07)	0.05 (0.06)
DPVC	**0.57** (0.13)	0.66 (0.07)	0.64 (0.14)	0.87 (0.17)	NA	NA
$\Sigma_j \sim \text{Uniform}_{d_j}$, no noise						
Proposed ($\beta = 0.01$)	0.15 (0.03)	0.33 (0.03)	0.87 (0.1)	0.98 (0.03)	**1.0** (0.0)	0.98 (0.03)
Proposed ($\beta = 0.02$)	0.15 (0.03)	0.33 (0.03)	0.87 (0.1)	0.97 (0.04)	**1.0** (0.0)	0.97 (0.04)
Proposed ($\beta = 0.03$)	0.16 (0.03)	0.31 (0.03)	0.67 (0.18)	0.97 (0.04)	0.98 (0.03)	0.97 (0.04)
Basic inverse Wishart prior	**0.17** (0.05)	0.33 (0.02)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
EBIC ($\gamma = 0$)	0.08 (0.09)	**0.6** (0.23)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
EBIC ($\gamma = 0.5$)	0.16 (0.03)	0.33 (0.04)	0.98 (0.03)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
EBIC ($\gamma = 1.0$)	0.16 (0.03)	0.31 (0.03)	0.91 (0.12)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
AIC	0.08 (0.08)	0.52 (0.33)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
Calinski–Harabasz Index	0.16 (0.06)	0.53 (0.3)	0.64 (0.15)	0.63 (0.28)	0.88 (0.17)	0.96 (0.08)
CGL (ALC)	0.0 (0.01)	0.0 (0.0)	0.0 (0.01)	0.15 (0.16)	0.15 (0.21)	0.12 (0.06)
DPVC	0.02 (0.01)	0.0 (0.04)	0.23 (0.14)	0.25 (0.13)	NA	NA

Ground truth is 4 unbalanced clusters with sizes 20, 10, 5, 5. Shows the ANMI of the selected models (standard deviation in brackets). No noise is added
The best ANMI scores are highlighted in bold

practice,[4] we always add 0.001 times the identity matrix to the covariance matrix (see also Ledoit and Wolf (2004)).

Finally, we also compare the proposed method to two previous approaches for variable clustering: the Clustered Graphical Lasso (CGL) as proposed in (Tan et al. 2015), and the Dirichlet process variable clustering (DPVC) model as proposed in (Palla et al. 2012), for which the implementation is available. DPVC models the number of clusters using a Dirichlet process. CGL uses for model selection the mean squared error for recovering randomly left-out elements of the covariance matrix. CGL uses for clustering either the single linkage clustering (SLC) or the average linkage clustering (ALC) method. For conciseness, we show only the results for ALC, since they tended to be better than SLC.

[4] In particular for the mutual funds data in the next section, where the covariance matrix was bad conditioned.

Table 6 Evaluation of clustering results with $d = 40$, $n \in \{20, 40, 400, 4000, 40{,}000, 4{,}000{,}000\}$

	20	40	400	4000	40,000	4,000,000
$\Sigma_j \sim \text{InvW}(d_j + 1, I_{d_j})$, $\Sigma_\epsilon \sim \text{InvW}(d + 1, I_d)$, $\eta = 0.01$						
Proposed ($\beta = 0.01$)	0.45 (0.14)	0.75 (0.15)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
Proposed ($\beta = 0.02$)	0.39 (0.09)	0.75 (0.15)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	0.98 (0.03)
Proposed ($\beta = 0.03$)	0.39 (0.09)	0.7 (0.18)	**1.0** (0.0)	0.97 (0.06)	**1.0** (0.0)	0.98 (0.03)
Basic inverse Wishart prior	**0.48** (0.15)	**0.8** (0.09)	**1.0** (0.0)	0.91 (0.11)	0.39 (0.13)	0.42 (0.12)
EBIC ($\gamma = 0$)	0.12 (0.08)	0.67 (0.12)	**1.0** (0.0)	0.91 (0.11)	0.48 (0.17)	0.42 (0.12)
EBIC ($\gamma = 0.5$)	0.19 (0.08)	0.32 (0.04)	0.97 (0.03)	**1.0** (0.0)	0.54 (0.26)	0.42 (0.12)
EBIC ($\gamma = 1.0$)	0.17 (0.07)	0.28 (0.07)	0.96 (0.03)	**1.0** (0.0)	0.68 (0.24)	0.42 (0.12)
AIC	0.06 (0.09)	0.3 (0.34)	**1.0** (0.0)	0.4 (0.1)	0.39 (0.13)	0.42 (0.12)
Calinski–Harabasz Index	0.2 (0.06)	0.13 (0.2)	0.45 (0.27)	0.59 (0.17)	0.7 (0.21)	0.77 (0.03)
CGL (ALC)	0.08 (0.06)	0.05 (0.03)	0.04 (0.03)	0.03 (0.02)	0.03 (0.02)	0.04 (0.04)
DPVC	0.28 (0.04)	0.35 (0.07)	0.57 (0.08)	0.4 (0.12)	NA	NA
$\Sigma_j \sim \text{InvW}(d_j + 1, I_{d_j})$, $\Sigma_\epsilon \sim \text{InvW}(d + 1, I_d)$, $\eta = 0.1$						
Proposed ($\beta = 0.01$)	0.09 (0.11)	**0.42** (0.12)	0.84 (0.1)	0.42 (0.16)	0.18 (0.22)	0.24 (0.18)
Proposed ($\beta = 0.02$)	0.09 (0.11)	**0.42** (0.13)	0.88 (0.11)	0.85 (0.15)	**0.99** (0.02)	**0.9** (0.09)
Proposed ($\beta = 0.03$)	**0.15** (0.06)	**0.42** (0.13)	**0.89** (0.09)	**0.92** (0.07)	**0.99** (0.02)	**0.9** (0.09)
Basic inverse Wishart prior	0.11 (0.14)	**0.42** (0.13)	0.84 (0.1)	0.2 (0.2)	0.0 (0.01)	0.1 (0.17)
EBIC ($\gamma = 0$)	0.04 (0.05)	0.24 (0.06)	0.88 (0.11)	0.2 (0.2)	0.0 (0.01)	0.1 (0.17)
EBIC ($\gamma = 0.5$)	0.05 (0.02)	0.19 (0.04)	0.74 (0.19)	0.44 (0.17)	0.0 (0.01)	0.1 (0.17)
EBIC ($\gamma = 1.0$)	0.05 (0.02)	0.19 (0.04)	0.41 (0.06)	0.78 (0.12)	0.0 (0.01)	0.1 (0.17)
AIC	0.0 (0.01)	0.15 (0.21)	0.19 (0.2)	0.2 (0.2)	0.0 (0.01)	0.1 (0.17)
Calinski–Harabasz Index	0.06 (0.03)	0.17 (0.11)	0.68 (0.25)	0.67 (0.2)	0.83 (0.17)	0.76 (0.04)
CGL (ALC)	0.04 (0.04)	0.03 (0.02)	0.05 (0.06)	0.1 (0.11)	0.05 (0.07)	0.08 (0.09)
DPVC	0.13 (0.05)	0.16 (0.05)	0.3 (0.13)	0.07 (0.03)	NA	NA
$\Sigma_j \sim \text{Uniform}_{d_j}$, $\Sigma_\epsilon \sim \text{Uniform}_d$, $\eta = 0.01$						
Proposed ($\beta = 0.01$)	0.11 (0.02)	0.32 (0.04)	0.74 (0.15)	**0.83** (0.1)	**0.59** (0.32)	0.5 (0.33)
Proposed ($\beta = 0.02$)	0.11 (0.02)	0.32 (0.04)	0.61 (0.17)	**0.83** (0.1)	**0.59** (0.32)	**0.59** (0.32)
Proposed ($\beta = 0.03$)	0.11 (0.02)	0.32 (0.04)	0.43 (0.06)	**0.83** (0.1)	**0.59** (0.32)	**0.59** (0.32)
Basic inverse Wishart prior	0.11 (0.02)	0.32 (0.04)	**0.84** (0.05)	0.28 (0.0)	0.11 (0.14)	0.17 (0.23)
EBIC ($\gamma = 0$)	**0.18** (0.13)	**0.43** (0.05)	0.76 (0.13)	0.22 (0.12)	0.11 (0.14)	0.06 (0.11)
EBIC ($\gamma = 0.5$)	0.11 (0.02)	0.32 (0.04)	**0.84** (0.05)	0.51 (0.3)	0.11 (0.14)	0.06 (0.11)
EBIC ($\gamma = 1.0$)	0.11 (0.02)	0.32 (0.04)	0.79 (0.13)	0.67 (0.24)	0.11 (0.14)	0.06 (0.11)
AIC	0.14 (0.05)	0.16 (0.28)	0.17 (0.23)	0.22 (0.12)	0.09 (0.12)	0.06 (0.11)
Calinski–Harabasz Index	0.14 (0.08)	0.32 (0.3)	0.34 (0.33)	0.68 (0.22)	0.25 (0.27)	0.41 (0.32)
CGL (ALC)	0.0 (0.0)	0.0 (0.0)	0.01 (0.04)	0.0 (0.01)	0.02 (0.02)	0.01 (0.01)
DPVC	0.01 (0.01)	0.03 (0.06)	0.2 (0.05)	0.01 (0.02)	NA	NA
$\Sigma_j \sim \text{Uniform}_{d_j}$, $\Sigma_\epsilon \sim \text{Uniform}_d$, $\eta = 0.1$						
Proposed ($\beta = 0.01$)	0.1 (0.02)	0.34 (0.07)	0.68 (0.18)	0.6 (0.31)	0.09 (0.12)	0.06 (0.11)
Proposed ($\beta = 0.02$)	0.11 (0.02)	0.34 (0.07)	0.65 (0.21)	**0.7** (0.13)	0.21 (0.21)	0.28 (0.26)
Proposed ($\beta = 0.03$)	0.11 (0.02)	0.32 (0.06)	0.58 (0.2)	**0.7** (0.13)	**0.32** (0.22)	0.28 (0.26)
Basic inverse Wishart prior	0.14 (0.03)	0.37 (0.08)	**0.78** (0.1)	0.0 (0.02)	0.09 (0.12)	0.06 (0.11)
EBIC ($\gamma = 0$)	**0.16** (0.05)	**0.49** (0.21)	0.71 (0.14)	0.0 (0.02)	0.09 (0.12)	0.06 (0.11)
EBIC ($\gamma = 0.5$)	0.11 (0.01)	0.36 (0.08)	0.77 (0.13)	0.06 (0.11)	0.09 (0.12)	0.06 (0.11)
EBIC ($\gamma = 1.0$)	0.11 (0.01)	0.31 (0.05)	0.7 (0.16)	0.12 (0.14)	0.09 (0.12)	0.06 (0.11)
AIC	0.15 (0.05)	0.05 (0.12)	0.06 (0.11)	0.0 (0.02)	0.09 (0.12)	0.06 (0.11)

Table 6 continued

	20	40	400	4000	40,000	4,000,000
Calinski–Harabasz Index	**0.16** (0.05)	0.29 (0.26)	0.42 (0.23)	0.45 (0.38)	0.09 (0.12)	**0.33** (0.31)
CGL (ALC)	0.0 (0.01)	0.0 (0.01)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.01)
DPVC	0.0 (0.04)	0.03 (0.05)	0.11 (0.13)	0.02 (0.03)	NA	NA

Ground truth is 4 unbalanced clusters with sizes 20, 10, 5, 5. Shows the ANMI of the selected models (standard deviation in brackets). Noise is added to the precision matrix
The best ANMI scores are highlighted in bold

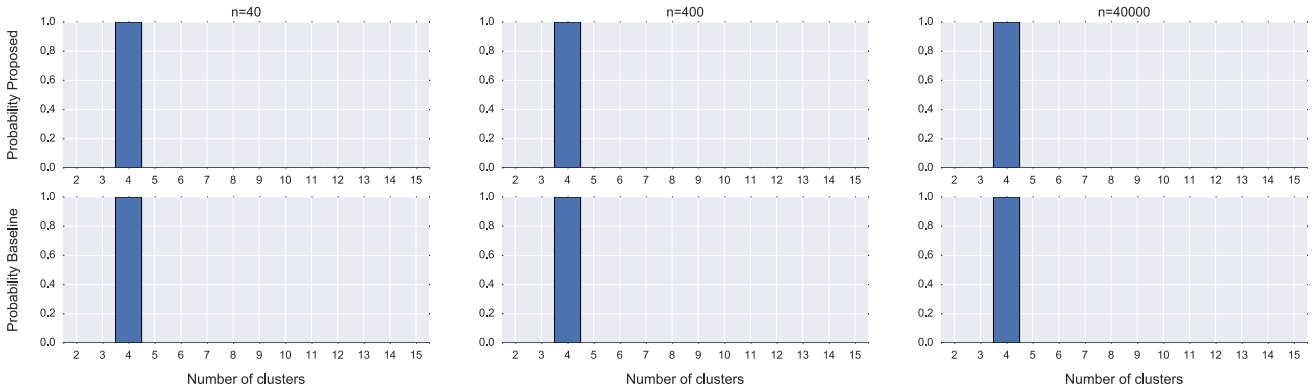

Fig. 3 Posterior distribution of the number of clusters of the proposed method (top row) and the basic inverse Wishart prior model (bottom row). Ground truth is 4 clusters; there is no noise on the precision matrix

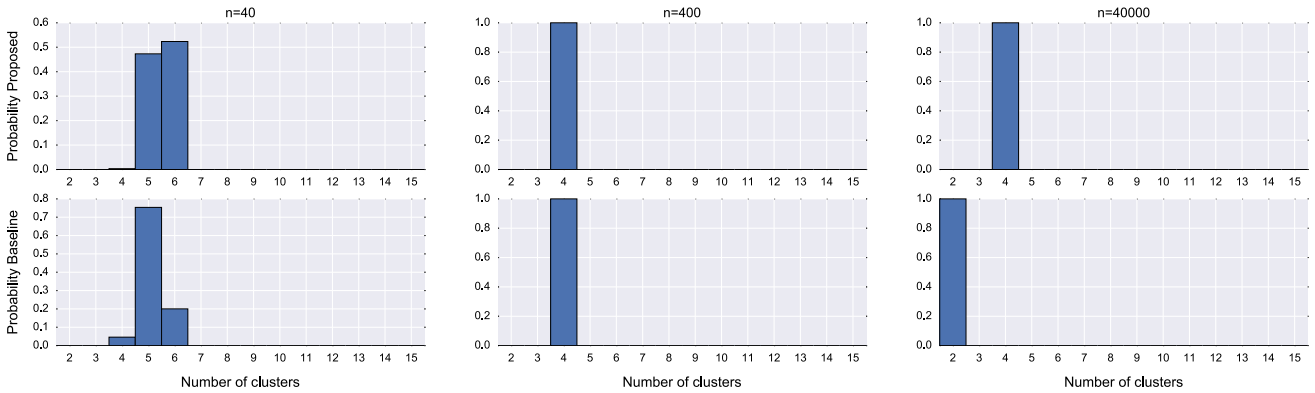

Fig. 4 Posterior distribution of the number of clusters of the proposed method (top row) and the basic inverse Wishart prior model (bottom row). Ground truth is 4 clusters; noise was added to the precision matrix

A summary of the experiments, with noise level $\eta \in \{0.0, 0.01, 0.1\}$, limited to the proposed method, EBIC, and Calinski–Harabasz Index, is shown in Figs. 1 and 2, for balanced and unbalanced clusters, respectively. Detailed results of all experiments are shown in Tables 3 and 4, for balanced clusters, and Tables 5 and 6, for unbalanced clusters. The tables also contain the performance of the proposed method for $\beta \in \{0, 0.01, 0.02, 0.03\}$. Note that $\beta = 0.0$ corresponds to the basic inverse Wishart prior model for which we can calculate the marginal likelihood analytically.

Comparing the proposed method with different β, we see that $\beta = 0.02$ offers good clustering performance in the no noise and noisy setting. In contrast, model selection with EBIC and AIC performs, as expected, well in the no noise scenario; however, in the noisy setting they tend to select incorrect clusterings. In particular, for large sample sizes EBIC tends to fail to identify correct clusterings.

The Calinski–Harabasz Index performs well in the noisy settings, whereas in the no noise setting it performs unsatisfactory.

In Figs. 3 and 4, we show the posterior distribution with and without noise on the precision matrix, respectively.[5] In both cases, given that the sample size n is large enough, the

[5] Same setting as before, $d = 40$, $\Sigma_j \sim \text{InvW}(d_j + 1, I_{d_j})$. Noise is $\Sigma_\epsilon \sim \text{InvW}(d + 1, I_d)$, $\eta = 0.01$. Proposed method $\beta = 0.02$.

Table 7 Comparison of variational and MCMC estimate. Evaluation of clustering results for $d = 12$, $n \in \{12, 120, 1200, 1{,}200{,}000\}$

	12	120	1200	1,200,000
$\Sigma_j \sim \text{InvW}(d_j + 1, I_{d_j})$, no noise				
Proposed, variational	**0.39** (0.23)	**0.89** (0.09)	0.96 (0.07)	0.82 (0.11)
Proposed, MCMC	0.37 (0.23)	**0.89** (0.09)	0.96 (0.07)	0.9 (0.14)
Basic inverse Wishart prior	**0.39** (0.23)	**0.89** (0.09)	**1.0** (0.0)	**1.0** (0.0)
$\Sigma_j \sim \text{Uniform}_{d_j}$, no noise				
Proposed, variational	**0.76** (0.17)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
Proposed, MCMC	0.66 (0.1)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
Basic inverse Wishart prior	**0.76** (0.17)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
$\Sigma_j \sim \text{InvW}(d_j + 1, I_{d_j})$, $\Sigma_\epsilon \sim \text{InvW}(d + 1, I_d)$, $\eta = 0.01$				
Proposed, variational	**0.42** (0.27)	0.8 (0.16)	**1.0** (0.0)	**0.96** (0.07)
Proposed, MCMC	0.17 (0.24)	0.8 (0.16)	**1.0** (0.0)	**0.96** (0.07)
Basic inverse Wishart prior	**0.42** (0.27)	**0.94** (0.12)	0.93 (0.13)	0.34 (0.04)
$\Sigma_j \sim \text{InvW}(d_j + 1, I_{d_j})$, $\Sigma_\epsilon \sim \text{InvW}(d + 1, I_d)$, $\eta = 0.1$				
Proposed, variational	0.11 (0.16)	0.57 (0.07)	0.55 (0.26)	**0.78** (0.2)
Proposed, MCMC	0.09 (0.06)	**0.61** (0.13)	**0.61** (0.23)	**0.78** (0.2)
Basic inverse Wishart prior	**0.16** (0.15)	0.54 (0.1)	0.28 (0.15)	0.21 (0.18)
$\Sigma_j \sim \text{Uniform}_{d_j}$, $\Sigma_\epsilon \sim \text{Uniform}_d$, $\eta = 0.01$				
Proposed, variational	0.79 (0.12)	0.82 (0.26)	0.73 (0.33)	**0.96** (0.07)
Proposed, MCMC	**0.82** (0.11)	**0.96** (0.09)	**0.75** (0.31)	**0.96** (0.07)
Basic inverse Wishart prior	0.79 (0.12)	0.48 (0.15)	0.28 (0.09)	0.28 (0.09)
$\Sigma_j \sim \text{Uniform}_{d_j}$, $\Sigma_\epsilon \sim \text{Uniform}_d$, $\eta = 0.1$				
Proposed, variational	0.67 (0.22)	**0.24** (0.24)	0.32 (0.0)	**0.35** (0.18)
Proposed, MCMC	0.68 (0.17)	**0.24** (0.24)	**0.46** (0.27)	**0.35** (0.18)
Basic inverse Wishart prior	**0.69** (0.21)	0.13 (0.11)	0.26 (0.13)	0.28 (0.09)

Ground truth is 4 balanced clusters. $\beta = 0.02$. Shows the ANMI of the selected models (standard deviation in brackets)
The best ANMI scores are highlighted in bold

proposed method is able to estimate correctly the number of clusters. In contrast, the basic inverse Wishart prior model underestimates the number of clusters for large n and existence of noise in the precision matrix.

5.3 Comparison of variational and MCMC estimate

Here, we compare our variational approximation with MCMC on a small scale simulated problem where it is computationally feasible to estimate the marginal likelihood with MCMC. We generated synthetic data as in the previous section, only with the difference that we set the number of variables d to 12.

The number of samples M for MCMC was set to 10,000, where we used 10% as burn-in. For two randomly picked clusterings for $n = 12$, and $n = 1{,}200{,}000$, we checked the acceptance rates and convergence using the multivariate extension of the Gelman–Rubin diagnostic (Brooks and Gelman 1998). The average acceptance rates were around 80%, and the potential scale reduction factor was 1.01.

The runtime of MCMC was around 40 minutes for evaluating one clustering, whereas for the variational approximation the runtime was around 2 seconds.[6] The results are shown in Table 7, suggesting that the quality of the selected clusterings using the variational approximation is similar to MCMC.

6 Real data experiments

In this section, we investigate the properties of the proposed model selection criterion on three real data sets. In all cases, we use the spectral clustering algorithm from "Appendix C" to create cluster candidates. All variables were normalized to have mean 0 and variance 1. For all methods, except DPVC, the number of clusters is considered to be in $\{2, 3, 4, \ldots, \min(p - 1, 15)\}$. DPVC automatically selects the number of clusters by assuming a Dirichlet process prior. We evaluated the proposed method with $\beta = 0.02$ using the variational approximation.

[6] Runtime on one core of Intel(R) Xeon(R) CPU 2.30GHz.

Table 8 Evaluation of selected clusterings of the mutual funds data

Proposed and EBIC ($\gamma = 0.0$) [number of clusters = 6, ANMI = 0.48]	
U.S. bond funds	2 2 2 2 2 2 2 4 2 2 2 2 2
U.S. stock funds	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 5 1 4 6
balanced funds	1 1 1 1 1 1 1
international stock funds	1 3 1 1 3 1 3 3 1
basic inverse Wishart prior [number of clusters = 3, ANMI = 0.42]	
U.S. bond funds	2 2 2 2 2 2 2 2 2 2 2 2 2
U.S. stock funds	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 1 1 1
balanced funds	1 1 1 1 1 1 1
international stock funds	1 1 1 1 1 1 1 1 1
EBIC ($\gamma = 0.5$) [number of clusters = 11, ANMI = 0.32]	
U.S. bond funds	2 9 2 9 2 2 2 1 10 9 2 2 2
U.S. stock funds	7 11 7 11 7 11 7 7 11 5 7 11 5 1 8 7 11 5 5 5 5 5 5 8 5 4 8 8 6
balanced funds	11 7 8 7 11 7 11
international stock funds	1 3 1 1 3 1 3 3 3
EBIC ($\gamma = 1.0$) [number of clusters = 14, ANMI = 0.25]	
U.S. bond funds	2 9 2 9 2 14 2 1 14 9 10 10 10
U.S. stock funds	12 8 12 6 12 8 12 12 8 6 12 8 6 3 11 6 8 5 7 5 5 5 5 6 11 5 11 15 4 11
balanced funds	8 12 1 12 8 6 7
international stock funds	3 13 3 3 13 3 13 13 13
AIC and Calinski-Harabaz Index [number of clusters = 2, ANMI = 0]	
U.S. bond funds	1 1 1 1 1 1 1 1 1 1 1 1 1
U.S. stock funds	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1
balanced funds	1 1 1 1 1 1 1
international stock funds	1 1 1 1 1 1 1 1 1
CGL (ALC) [number of clusters = 3, ANMI = 0.36]	
U.S. bond funds	1 1 1 1 1 1 1 3 1 1 1 1 1
U.S. stock funds	2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3
balanced funds	2 2 2 2 3 2 2
international stock funds	2 2 2 2 2 2 2 3 2
DPVC [number of clusters = 2, ANMI = 0.35]	
U.S. bond funds	1 1 1 1 1 1 2 1 1 1 1 1
U.S. stock funds	2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
balanced funds	2 2 2 2 2 2 2
international stock funds	2 2 2 2 2 2 2 2 2

Colors highlight the type of fund. Numbers denote the cluster id assigned by the respective method. Here, the size of the restricted hypotheses space $|\mathscr{C}^*|$ found by spectral clustering was 128

6.1 Mutual funds

Here, we use the mutual funds data, which has been previously analyzed in (Scott and Carvalho 2008; Marlin et al. 2009). The data contain 59 mutual funds (d =59) grouped into 4 clusters: US bond funds, US stock funds, balanced funds (containing US stocks and bonds), and international stock funds. The number of observations is 86.

The results of all methods are visualized in Table 8. It is difficult to interpret the results produced by EBIC ($\gamma = 1.0$), AIC, and the Calinski–Harabasz Index. In contrast, the proposed method and EBIC ($\gamma = 0.0$) produce results that are easier to interpret. In particular, our results suggest that there is a considerable correlation between the balanced funds and the US stock funds which was also observed in Marlin et al. (2009).

In Fig. 5, we show a two-dimensional representation of the data, that was found using Laplacian eigenmaps (Belkin and Niyogi 2003). The figure supports the claim that balanced funds and the US stock funds have similar behavior.

6.2 Gene regulations

We tested our method also on the gene expression data that was analyzed in (Hirose et al. 2017). The data consist of 11 genes with 445 gene expressions. The true gene reg-

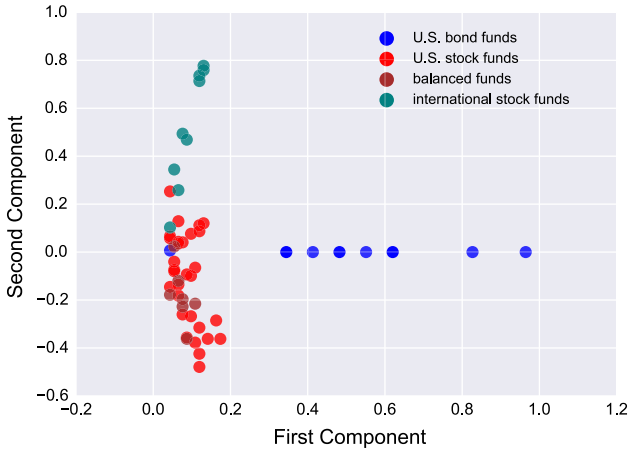

Fig. 5 Two-dimensional representation of the mutual funds data suggesting that balanced funds and US stock funds are difficult to separate (one cluster), whereas US bond funds and international stock funds appear to form mostly separate clusters

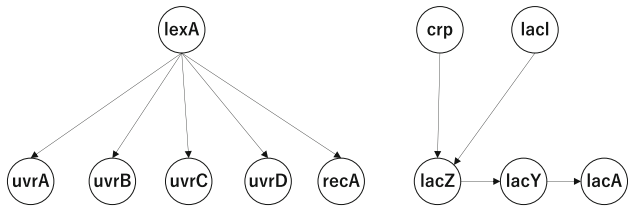

Fig. 6 Gene regulations of *E. coli* as given in (Hirose et al. 2017; Albersts et al. 2014) suggesting that the gene groups {lexA, uvrA, uvrB, uvrC, uvrD, recA} and {crp, lacI, lacZ, lacY, lacA} should be separated

ularizations are known in this case and shown in Fig. 6, adapted from (Hirose et al. 2017). The most important fact is that there are two independent groups of genes and any clustering that mixes these two can be considered as wrong.

We show the results of all methods in Fig. 7, where we mark each cluster with a different color superimposed on the true regularization structure. Here, only the clustering selected by the proposed method, EBIC ($\gamma = 1.0$) and Calinski–Harabasz correctly divides the two group of genes.

6.3 Aviation sensors

As a third data set, we use the flight aviation data set from NASA.[7] The data set contains sensor information sampled from airplanes during operation. We extracted the information of 16 continuous-valued sensors that were recorded for different flights with in total 25,032,364 samples.

The clustering results are shown in Table 9. The data set does not have any ground truth, but the clustering result of our proposed method is reasonable: Cluster 9 groups sensors that measure or affect altitude,[8] Cluster 8 correctly clusters the left and right sensors for measuring the rotation around the axis pointing through the noise of the aircraft, in Cluster 2 all sensors that measure the angle between chord and flight direction are grouped together. It also appears reasonable that the yellow hydraulic system of the left part of the plane has little direct interaction with the green hydraulic system of the right part (Cluster 1 and Cluster 4). And the sensor for the rudder, influencing the direction of the plane, is mostly independent of the other sensors (Cluster 5).

In contrast, the clustering selected by the basic inverse Wishart prior, EBIC, and AIC is difficult to interpret. We note that we did not compare to DPVC, since the large number of samples made the MCMC algorithm of DPVC infeasible.

7 Discussion and conclusions

We have introduced a new method for evaluating variable clusterings based on the marginal likelihood of a Bayesian model that takes into account noise on the precision matrix. Since the calculation of the marginal likelihood is analytically intractable, we proposed two approximations: a variational approximation and an approximation based on MCMC. Experimentally, we found that the variational approximation is considerably faster than MCMC and also leads to accurate model selection.

We compared our proposed method to several standard model selection criteria. In particular, we compared to BIC and extended BIC (EBIC) which are often the method of choice for model selection in Gaussian graphical models. However, we emphasize that EBIC was designed to handle the situation where d is in the order of n, and has not been designed to handle noise. As a consequence, our experiments showed that in practice its performance depends highly on the choice of the γ parameter. In contrast, the proposed method, with fixed hyper-parameters, shows better performance on various simulated and real data.

We also compared our method to other two previously proposed methods, namely Cluster Graphical Lasso (CGL) (Tan et al. 2015) and Dirichlet Process Variable Clustering (DPVC) (Palla et al. 2012) that performs jointly clustering

[7] https://c3.nasa.gov/dashlink/projects/85/ where we use all records from Tail 687.

[8] The elevator position of an airplane influences the altitude, and the static pressure system of an airplane measures the altitude.

Fig. 7 Clusterings of gene regulations network of E. coli. The clustering results are visualized by different colors. Here, the size of the restricted hypotheses space $|\mathscr{G}^*|$ found by spectral clustering was 18. Only the proposed method, EBIC ($\gamma = 1.0$), and Calinski–Harabasz correctly divide the gene groups {lexA, uvrA, uvrB, uvrC, uvrD, recA} and {crp, lacI, lacZ, lacY, lacA}

and model selection. However, it appears that in many situations the model selection algorithm of CGL is not able to detect the true model, even if there is no noise. On the other hand, the Dirichlet process assumption by DPVC appears to be very restrictive, leading again to many situations where the true model (clustering) is missed. Overall, our method performs better in terms of selecting the correct clustering on synthetic data with ground truth, and selects meaningful clusters on real data.

The python source code for variable clustering and model selection with the proposed method and all baselines is available at https://github.com/andrade-stats/robustBayesClustering.

Table 9 Evaluation of selected clusterings of the Aviation Sensor Data with 16 variables

Proposed	
Cluster 1	BRAKE PRESSURE LH YELLOW
Cluster 2	INDICATED ANGLE OF ATTACK, ANGLE OF ATTACK 2, ANGLE OF ATTACK 1
Cluster 3	ROLL SPOILER RIGHT
Cluster 4	BRAKE PRESSURE RH GREEN
Cluster 5	RUDDER POSITION
Cluster 6	AILERON POSITION RH, AILERON POSITION LH
Cluster 7	ROLL SPOILER LEFT
Cluster 8	PITCH TRIM POSITION
Cluster 9	STATIC PRESSURE LSP, TOTAL PRESSURE LSP, AVARAGE STATIC PRESSURE LSP, ELEVATOR POSITION LEFT,ELEVATOR POSITION RIGHT
Basic inverse Wishart prior, EBIC ($\gamma \in \{0.0, 0.5, 1.0\}$), AIC	
Cluster 1	STATIC PRESSURE LSP, INDICATED ANGLE OF ATTACK, TOTAL PRESSURE LSP, RUDDER POSITION, AILERON POSITION RH, AVARAGE STATIC PRESSURE LSP, ELEVATOR POSITION LEFT, ELEVATOR POSITION RIGHT, PITCH TRIM POSITION, ANGLE OF ATTACK 2, ANGLE OF ATTACK 1, AILERON POSITION LH, ROLL SPOILER LEFT, BRAKE PRESSURE LH YELLOW, ROLL SPOILER RIGHT
Cluster 2	BRAKE PRESSURE RH GREEN
Calinski–Harabasz Index	
Cluster 1	STATIC PRESSURE LSP, TOTAL PRESSURE LSP, AILERON POSITION RH, AVARAGE STATIC PRESSURE LSP, ELEVATOR POSITION LEFT, ELEVATOR POSITION RIGHT, BRAKE PRESSURE RH GREEN, AILERON POSITION LH, BRAKE PRESSURE LH YELLOW
Cluster 2	INDICATED ANGLE OF ATTACK, ANGLE OF ATTACK 2, ANGLE OF ATTACK 1
Cluster 3	RUDDER POSITION, PITCH TRIM POSITION, ROLL SPOILER LEFT, ROLL SPOILER RIGHT
CGL (ALC)	
Cluster 1	STATIC PRESSURE LSP, TOTAL PRESSURE LSP, AVARAGE STATIC PRESSURE LSP, ELEVATOR POSITION LEFT, ELEVATOR POSITION RIGHT, BRAKE PRESSURE LH YELLOW
Cluster 2	INDICATED ANGLE OF ATTACK, RUDDER POSITION, AILERON POSITION RH, PITCH TRIM POSITION, BRAKE PRESSURE RH GREEN, ANGLE OF ATTACK 2, ANGLE OF ATTACK 1, AILERON POSITION LH, ROLL SPOILER LEFT, ROLL SPOILER RIGHT

Here, the size of the restricted hypotheses space $|\mathscr{C}^*|$ found by spectral clustering was 28

Appendix A Convergence of 3-block ADMM

We can write the optimization problem in (4) as

$$\begin{aligned}&\text{minimize } f_1(X_\epsilon) + f_2(X_1, \ldots, X_k) + f_3(Z)\\ &\text{subject to}\\ &\quad -X - \beta X_\epsilon + Z = 0,\\ &X_\epsilon, X_1, \ldots, X_k \succ 0,\end{aligned}$$

with

$$\begin{aligned}f_1(X_\epsilon) &:= \text{trace}(A_\epsilon X_\epsilon) - a_\epsilon \cdot \log|X_\epsilon|,\\ f_2(X_1, \ldots, X_k) &:= \sum_{j=1}^{k} \Big(\text{trace}(A_j X_j) - a_j \cdot \log|X_j|\Big),\\ f_3(Z) &:= n \cdot \text{trace}(SZ) - n \cdot \log|Z|.\end{aligned}$$

First note that the functions f_1, f_2, and f_3 are convex proper closed functions. Since $X_\epsilon, X_1, \ldots, X_k \succ 0$, we have due to the equality constraint that $Z \succ 0$. Assuming that the global minima is attained, we can assume that $Z \preceq \sigma I$, for some large enough $\sigma > 0$. As a consequence, we have that $\nabla^2 f_3(Z) = Z^{-1} \otimes Z^{-1} \succeq \sigma^{-2} I$, and therefore f_3 is a strongly convex function. Analogously, we have that f_1 and f_2 are strongly convex functions, and therefore also coercive. This allows us to apply Theorem 3.2 in (Lin et al. 2018) which guarantees the convergence of the 3-block ADMM.

Appendix B Derivation of variational approximation

Here, we give more details of the KL-divergence minimization from Sect. 4.2.2. Recall, that the remaining parameters $\nu_{g,\epsilon} \in \mathbb{R}$ and $\nu_{g,j} \in \mathbb{R}$ are optimized by minimizing the KL-divergence between the factorized distribution g and the posterior distribution $p(\Sigma_\epsilon, \Sigma_1, \ldots \Sigma_k | \mathbf{x}_1, \ldots, \mathbf{x}_n, \boldsymbol{\eta}, \mathcal{C})$. We have

$$\begin{aligned}KL(g||p) &= -\int g_\epsilon(\Sigma_\epsilon) \cdot \prod_{j=1}^{k} g_j(\Sigma_j)\\ &\qquad \log \frac{p(\Sigma_\epsilon, \Sigma_1, \ldots \Sigma_k, \mathbf{x}_1, \ldots, \mathbf{x}_n | \boldsymbol{\eta}, \mathcal{C})}{g_\epsilon(\Sigma_\epsilon) \cdot \prod_{j=1}^{k} g_j(\Sigma_j)} d\Sigma_\epsilon d\Sigma + c\\ &= -\frac{1}{2}\mathbb{E}_{g_J, g_\epsilon}[n \cdot \log|(\Sigma^{-1} + \beta\Sigma_\epsilon^{-1})|]\\ &\quad - \frac{1}{2}\mathbb{E}_{g_\epsilon}[(\nu_\epsilon + d + 1) \cdot \log|\Sigma_\epsilon^{-1}|\\ &\quad - \text{trace}((\Sigma_{\epsilon,0} + \beta n S)\Sigma_\epsilon^{-1})] - \text{Entropy}[g_\epsilon]\\ &\quad + \sum_{j=1}^{k}\Big(-\frac{1}{2}\mathbb{E}_{g_j}[(\nu_j + d_j + 1) \cdot \log|\Sigma_j^{-1}|\\ &\quad -\text{trace}((\Sigma_{j,0} + nS_j)\Sigma_j^{-1})] - \text{Entropy}[g_j]\Big) + c\\ &= -\frac{1}{2} n\,\mathbb{E}_{g_J, g_\epsilon}[\log|\Sigma^{-1} + \beta\Sigma_\epsilon^{-1}|]\\ &\quad + \frac{1}{2}(\nu_\epsilon + d + 1)\,\mathbb{E}_{g_\epsilon}[\log|\Sigma_\epsilon|]\\ &\quad + \frac{1}{2}\text{trace}((\Sigma_{\epsilon,0} + \beta n S)\,\mathbb{E}_{g_\epsilon}[\Sigma_\epsilon^{-1}]) - \text{Entropy}[g_\epsilon]\\ &\quad + \frac{1}{2}\sum_{j=1}^{k}(\nu_j + d_j + 1)\,\mathbb{E}_{g_j}[\log|\Sigma_j|]\\ &\quad + \frac{1}{2}\sum_{j=1}^{k}\text{trace}((\Sigma_{j,0} + nS_j)\,\mathbb{E}_{g_j}[\Sigma_j^{-1}])\\ &\quad - \sum_{j=1}^{k}\text{Entropy}[g_j] + c,\end{aligned}$$

where c is a constant with respect to g_ϵ and g_j. However, the term $E_{g_J, g_\epsilon}[\log|\Sigma^{-1} + \beta\Sigma_\epsilon^{-1}|]$ cannot be solved analytically; therefore, we need to resort to some sort of approximation. Assuming that

$$E_{g_J, g_\epsilon}[\log|\Sigma^{-1} + \beta\Sigma_\epsilon^{-1}|] \approx E_{g_J, g_\epsilon}[\log|\Sigma^{-1}|],$$

we get

$$\begin{aligned}KL(g||p) \approx &-\frac{1}{2} n\,\mathbb{E}_{g_J, g_\epsilon}[\log|\Sigma^{-1}|]\\ &+ \frac{1}{2}(\nu_\epsilon + d + 1)\,\mathbb{E}_{g_\epsilon}[\log|\Sigma_\epsilon|]\\ &+ \frac{1}{2}\text{trace}((\Sigma_{\epsilon,0} + \beta n S)\,\mathbb{E}_{g_\epsilon}[\Sigma_\epsilon^{-1}]) - \text{Entropy}[g_\epsilon]\\ &+ \frac{1}{2}\sum_{j=1}^{k}(\nu_j + d_j + 1)\,\mathbb{E}_{g_j}[\log|\Sigma_j|]\\ &+ \frac{1}{2}\sum_{j=1}^{k}\text{trace}((\Sigma_{j,0} + nS_j)\,\mathbb{E}_{g_j}[\Sigma_j^{-1}])\\ &- \sum_{j=1}^{k}\text{Entropy}[g_j] + c\\ = &-\mathbb{E}_{g_\epsilon}\Big[\log\Big(|\Sigma_\epsilon|^{-\frac{1}{2}(\nu_\epsilon + d + 1)}\\ &\quad e^{-\frac{1}{2}\text{trace}((\Sigma_{\epsilon,0} + \beta n S)\Sigma_\epsilon^{-1})}\Big)\Big]\\ &- \text{Entropy}[g_\epsilon] - \sum_{j=1}^{k}\mathbb{E}_{g_j}\Big[\log\Big(|\Sigma_j|^{-\frac{1}{2}(\nu_j + n + d_j + 1)}\\ &\quad e^{-\frac{1}{2}\text{trace}((\Sigma_{j,0} + nS_j)\Sigma_j^{-1})}\Big)\Big] + \text{Entropy}[g_j] + c\end{aligned}$$

$$
\begin{aligned}
&= -\mathbb{E}_{g_\epsilon}[\log \mathrm{InvW}(\nu_\epsilon, \Sigma_{\epsilon,0} + \beta n S)] \\
&\quad - \mathrm{Entropy}[g_\epsilon] \\
&\quad - \sum_{j=1}^{k} \mathbb{E}_{g_j}[\log \mathrm{InvW}(\nu_j + n, \Sigma_{j,0} + n S_j)] \\
&\quad + \mathrm{Entropy}[g_j] + c' \\
&= KL(g_\epsilon \,||\, \mathrm{InvW}(\nu_\epsilon, \Sigma_{\epsilon,0} + \beta n S)) \\
&\quad + \sum_{j=1}^{k} KL(g_j \,||\, \mathrm{InvW}(\nu_j + n, \Sigma_{j,0} + n S_j)) \\
&\quad + c',
\end{aligned}
$$

where we used that $\mathbb{E}_{g_J, g_\epsilon}[\log|\Sigma^{-1}|] = -\sum_{j=1}^{k} \mathbb{E}_{g_j}[\log|\Sigma_j|]$, and c' is a constant with respect to g_ϵ and g_j.

From the above expression, we see that we can optimize the parameters of g_ϵ and g_j independently from each other. The optimal parameter $\hat{\nu}_{g,\epsilon}$ for g_ϵ is

$$
\begin{aligned}
\hat{\nu}_{g,\epsilon} &= \arg\min_{\nu_{g,\epsilon}} KL(g_\epsilon \,||\, \mathrm{InvW}(\nu_\epsilon, \Sigma_{\epsilon,0} + \beta n S)) \\
&= \arg\min_{\nu_{g,\epsilon}} (\nu_\epsilon + d + 1)\, \mathbb{E}_{g_\epsilon}[\log|\Sigma_\epsilon|] \\
&\quad + \mathrm{trace}((\Sigma_{\epsilon,0} + \beta n S)\, \mathbb{E}_{g_\epsilon}[\Sigma_\epsilon^{-1}]) - 2 \cdot \mathrm{Entropy}[g_\epsilon] \\
&= \arg\min_{\nu_{g,\epsilon}} (\nu_\epsilon + d + 1) \left(-d \log 2 + d \log(\nu_{g,\epsilon} + d + 1) \right. \\
&\quad \left. + \log|\hat{\Sigma}_\epsilon| - \sum_{i=1}^{d} \psi\left(\frac{\nu_{g,\epsilon} - d + i}{2} \right) \right) \\
&\quad + \frac{\nu_{g,\epsilon}}{\nu_{g,\epsilon} + d + 1} \mathrm{trace}((\Sigma_{\epsilon,0} + \beta n S)\hat{\Sigma}_\epsilon^{-1}) \\
&\quad - 2 \log \Gamma_d\left(\frac{\nu_{g,\epsilon}}{2}\right) - \nu_{g,\epsilon} d - d(d+1) \log(\nu_{g,\epsilon} + d + 1) \\
&\quad + (\nu_{g,\epsilon} + d + 1) \sum_{i=1}^{d} \psi\left(\frac{\nu_{g,\epsilon} - d + i}{2} \right) \\
&= \arg\min_{\nu_{g,\epsilon}} p(\nu_\epsilon + d + 1) \log(\nu_{g,\epsilon} + d + 1) \\
&\quad - (\nu_\epsilon + d + 1) \sum_{i=1}^{d} \psi\left(\frac{\nu_{g,\epsilon} - d + i}{2} \right) \\
&\quad + \frac{\nu_{g,\epsilon}}{\nu_{g,\epsilon} + d + 1} \mathrm{trace}((\Sigma_{\epsilon,0} + \beta n S)\hat{\Sigma}_\epsilon^{-1}) \\
&\quad - 2 \log \Gamma_d\left(\frac{\nu_{g,\epsilon}}{2}\right) - \nu_{g,\epsilon} d - d(d+1) \log(\nu_{g,\epsilon} + d + 1) \\
&\quad + (\nu_{g,\epsilon} + d + 1) \sum_{i=1}^{d} \psi\left(\frac{\nu_{g,\epsilon} - d + i}{2} \right) \\
&= \arg\min_{\nu_{g,\epsilon}} \frac{\nu_{g,\epsilon}}{\nu_{g,\epsilon} + d + 1} \mathrm{trace}((\Sigma_{\epsilon,0} + \beta n S)\hat{\Sigma}_\epsilon^{-1}) \\
&\quad - 2 \log \Gamma_d\left(\frac{\nu_{g,\epsilon}}{2}\right) - \nu_{g,\epsilon} d + d \nu_\epsilon \log(\nu_{g,\epsilon} + d + 1) \\
&\quad + (\nu_{g,\epsilon} - \nu_\epsilon) \sum_{i=1}^{d} \psi\left(\frac{\nu_{g,\epsilon} - d + i}{2} \right).
\end{aligned}
$$

And analogously, we have

$$
\begin{aligned}
\hat{\nu}_{g,j} &= \arg\min_{\nu_{g,j}} \frac{\nu_{g,j}}{\nu_{g,j} + d_j + 1} \mathrm{trace}((\Sigma_{j,0} + n S_j)\hat{\Sigma}_j^{-1}) \\
&\quad - 2 \log \Gamma_{d_j}\left(\frac{\nu_{g,j}}{2}\right) - \nu_{g,j} d_j \\
&\quad + d_j(\nu_j + n) \log(\nu_{g,j} + d_j + 1) \\
&\quad + (\nu_{g,j} - \nu_j - n) \sum_{i=1}^{d_j} \psi\left(\frac{\nu_{g,j} - d_j + i}{2} \right).
\end{aligned}
$$

Appendix C Spectral clustering for variable clustering with the Gaussian graphical model

Let $S \in \mathbb{R}^{d \times d}$ denote the sample covariance matrix of the observed variables. Under the assumption that the observations are drawn i.i.d. from a multivariate normal distribution, with mean $\mathbf{0}$ and precision matrix $X + \beta X_\epsilon$, the log-likelihood[9] of the data is given by

$$
\frac{n}{2}(\log|X + \beta X_\epsilon| - \mathrm{trace}((X + \beta X_\epsilon)S)),
$$

where n is the number of observations. We assume that X is block sparse, i.e., a permutation matrix P exists such that $P^T X P$ is block diagonal. If we knew the number of blocks k, then we could estimate the block matrix X (and thus the variable clustering) by the following optimization problem.

Optimization Problem 1

$$
\begin{aligned}
&\underset{X \succ 0}{\text{minimize}} \; -\log|X + \beta X_\epsilon| + \mathrm{trace}((X + \beta X_\epsilon)S) \\
&\text{subject to} \\
&\text{X is block sparse with exactly } k \text{ blocks},
\end{aligned}
$$

where βX_ϵ is assumed to be a constant matrix with small entries. We claim that this can be reformulated, for any $q > 0$, as following.

Optimization Problem 2

$$
\begin{aligned}
&\underset{X \succ 0}{\text{minimize}} \; -\log|X + \beta X_\epsilon| + \mathrm{trace}((X + \beta X_\epsilon)S) \\
&\text{subject to} \\
&L_{ii} = \sum_{k \neq i} |X_{ik}|^q, \\
&L_{ij} = -|X_{ij}|^q \quad \text{for } i \neq j, \\
&rank(L) = p - k.
\end{aligned}
$$

[9] Up to a constant that does not depend on X.

Proposition 1 *Optimization problems 1 and 2 have the same solution. Moreover, the k-dimensional null space of L can be chosen such that each basis vector is the indicator vector for one variable block of X.*

Proof First let us define the matrix $\tilde{X}$, by $\tilde{X}_{ij} := |X_{ij}|^q$. Then clearly, iff X is block sparse with k blocks, so is $\tilde{X}$. Furthermore, $\tilde{X}_{ij} \geq 0$, and L is the unnormalized Laplacian as defined in (Von Luxburg 2007). We can therefore apply Proposition (2) of (Von Luxburg 2007), to find that the dimension of the eigenspace of L corresponding to eigenvalue 0, is exactly the number of blocks in $\tilde{X}$. Also from Proposition (2) of (Von Luxburg 2007) it follows that each such eigenvector $\mathbf{e}_k \in \mathbb{R}^d$ can be chosen such that it indicates the variables belonging to the same block, i.e., $\mathbf{e}_k(i) \neq 0$, iff variable i belongs to block k. □

Using the nuclear norm as a convex relaxation for the rank constraint, we have

$$\underset{X \succeq 0}{\text{minimize}} -\log|X + \beta X_\epsilon| + \text{trace}((X + \beta X_\epsilon)S) + \lambda_k ||L||_*$$

subject to

$$L_{ii} = \sum_{k \neq i} |X_{ik}|^q,$$

$$L_{ij} = -|X_{ij}|^q \quad \text{for } i \neq j.$$

with an appropriately chosen λ_k. By the definition of L, we have that L is positive semi-definite, and therefore $||L||_* = \text{trace}(L)$. As a consequence, we can rewrite the above problem as

$$X^* := \underset{X \succeq 0}{\arg\min} -\log|X + \beta X_\epsilon| + \text{trace}((X + \beta X_\epsilon)S) + \lambda_k \sum_{i \neq j} |X_{ij}|^q.$$

Finally, for the purpose of learning the Laplacian L, we ignore the term βX_ϵ and set it to zero. This will necessarily lead to an estimate of X^* that is not a clean block matrix, but has small nonzero entries between blocks. Nevertheless, spectral clustering is known to be robust to such violations (Ng et al. 2002). This leads to Algorithm 2 in Sect. 4.3.

References

Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: Parzen, E.K.G., Tanabe, K. (eds.) Reprint in Breakthroughs in Statistics, 1992, pp. 610–624. Springer, New York (1973)

Albersts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K., Walter, P.: Molecular Biology of the Cell: The Problems Book. Garland Science, New York (2014)

Anderson, T.W.: An Introduction to Multivariate Statistical Analysis, vol. 3. Wiley, New York (2004)

Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput. **15**(6), 1373–1396 (2003)

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3**(1), 1–122 (2011)

Brent, R.P.: Algorithms for finding zeros and extrema of functions without calculating derivatives. Technical report, Stanford University, Department of Computer Science (1971)

Brooks, S.P., Gelman, A.: General methods for monitoring convergence of iterative simulations. J. Comput. Graph. Stat. **7**(4), 434–455 (1998)

Caliński, T., Harabasz, J.: A dendrite method for cluster analysis. Commun. Stat. Theory Methods **3**(1), 1–27 (1974)

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A.: Stan: a probabilistic programming language. J. Stat. Softw. **76**(1), 1–32 (2017)

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

Chib, S.: Marginal likelihood from the Gibbs output. J. Am. Stat. Assoc. **90**(432), 1313–1321 (1995)

Chib, S., Jeliazkov, I.: Marginal likelihood from the Metropolis–Hastings output. J. Am. Stat. Assoc. **96**(453), 270–281 (2001)

Devijver, E., Gallopin, M.: Block-diagonal covariance selection for high-dimensional Gaussian graphical models. J. Am. Stat. Assoc. **113**(521), 306–314 (2018)

Foygel, R., Drton, M.: Extended Bayesian information criteria for Gaussian graphical models. In: Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R., Culotta, A. (eds.) Advances in Neural Information Processing Systems, pp. 604–612. Springer, New York (2010)

Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso. Biostatistics **9**(3), 432–441 (2008)

Hans, C., Dobra, A., West, M.: Shotgun stochastic search for "large p" regression. J. Am. Stat. Assoc. **102**(478), 507–516 (2007)

Hirose, K., Fujisawa, H., Sese, J.: Robust sparse Gaussian graphical modeling. J. Multivar. Anal. **161**, 172–190 (2017)

Hosseini, S.M.J., Lee, S.I.: Learning sparse Gaussian graphical models with overlapping blocks. In: Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems, pp. 3801–3809. MIT Press, Cambridge (2016)

Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)

Konishi, S., Ando, T., Imoto, S.: Bayesian information criteria and smoothing parameter selection in radial basis function networks. Biometrika **91**(1), 27–43 (2004)

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., Blei, D.M.: Automatic differentiation variational inference. J. Mach. Learn. Res. **18**(1), 430–474 (2017)

Ledoit, O., Wolf, M.: A well-conditioned estimator for large-dimensional covariance matrices. J. Multivar. Anal. **88**(2), 365–411 (2004)

Lenkoski, A., Dobra, A.: Computational aspects related to inference in Gaussian graphical models with the G-Wishart prior. J. Comput. Graph. Stat. **20**(1), 140–157 (2011)

Lin, T., Ma, S., Zhang, S.: Global convergence of unmodified 3-block ADMM for a class of convex minimization problems. J. Sci. Comput. **76**(1), 69–88 (2018)

Marlin, B.M., Murphy, K.P.: Sparse Gaussian graphical models with unknown block structure. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 705–712. ACM (2009)

Marlin, B.M., Schmidt, M., Murphy, K.P.: Group sparse priors for covariance estimation. In: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, pp. 383–392. AUAI Press (2009)

Ng, A.Y., Jordan, M.I., Weiss, Y.: Others: on spectral clustering: analysis and an algorithm. Adv. Neural Inf. Process. Syst. **2**, 849–856 (2002)

Palla, K., Ghahramani, Z., Knowles, D.A.: A nonparametric variable clustering model. In: Pereira, F., Burges, C., Bottou, L., Weinberger, K. (eds.) Advances in Neural Information Processing Systems, pp. 2987–2995. MIT Press, Cambridge (2012)

Ranganath, R., Gerrish, S., Blei, D.: Black box variational inference. In: Kaski, S., Corander, J. (eds.) Artificial Intelligence and Statistics, pp. 814–822. Springer, New York (2014)

Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**(2), 461–464 (1978)

Scott, J.G., Carvalho, C.M.: Feature-inclusion stochastic search for Gaussian graphical models. J. Comput. Graph. Stat. **17**(4), 790–808 (2008)

Sun, S., Zhu, Y., Xu, J.: Adaptive variable clustering in Gaussian graphical models. In: AISTATS, pp. 931–939 (2014)

Sun, S., Wang, H., Xu, J.: Inferring block structure of graphical models in exponential families. In: AISTATS (2015)

Tan, K.M., Witten, D., Shojaie, A.: The cluster graphical lasso for improved estimation of Gaussian graphical models. Comput. Stat. Data Anal. **85**, 23–36 (2015)

Vinh, N.X., Epps, J., Bailey, J.: Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. J. Mach. Learn. Res. **11**(Oct), 2837–2854 (2010)

Von Luxburg, U.: A tutorial on spectral clustering. Stat. Comput. **17**(4), 395–416 (2007)