Bayesian block-diagonal graphical models via the Fiedler prior

Julyan Arbel, Mario Beraha, Daria Bystrova

Bayesian block-diagonal graphical models via the Fiedler prior

Julyan Arbel[1] & Mario Beraha[2,3] & Daria Bystrova[1]

[1] *Univ. Grenoble Alpes, Inria, CNRS, LJK, 38000 Grenoble, France*
{julyan.arbel, daria.bystrova}@inria.fr
[2] *Department of Mathematics, Politecnico di Milano*, mario.beraha@polimi.it
[3] *Department of Computer Science, Universitá degli studi di Bologna*

Résumé. Nous étudions le problème de l'inférence de la structure d'indépendance conditionnelle entre les entrées d'un vecteur aléatoire gaussien, principalement dans le but d'obtenir des groupes de variables indépendantes. Cela peut se traduire par l'estimation d'une matrice de précision (inverse de la matrice de covariance) avec une structure bloc-diagonale. Cette approche se base sur des techniques de théorie spectrale des graphes et de clustering spectral. Nous proposons une nouvelle loi a priori, le prior de *Fiedler*, qui satisfait une propriété de *shrinkage* vers les matrices de précision à structure bloc-diagonale. Nous comparons le *shrinkage* induit par ce prior de Fiedler et par le Graphical Lasso, et comparons leurs performances sur un ensemble de données simulées.

Mots-clés. Modèles graphiques, matrice de précision, valeur de Fiedler, théorie spectrale des graphes.

Abstract. We study the problem of inferring the conditional independence structure between the entries of a Gaussian random vector. Our focus is on finding groups of independent variables. This can be translated into the estimation of a precision matrix (inverse of the covariance matrix) with a block-diagonal structure. We borrow ideas from spectral graph theory and spectral clustering and propose a novel prior called *Fiedler* prior showing shrinkage properties towards block-diagonal precision matrices. We compare the shrinkage induced by our prior and the popular Graphical Lasso prior, and compare their performance on a simulated dataset.

Keywords. Graphical models, Precision matrix, Fiedler value, Spectral graph theory.

1 Introduction

Understanding the dependence structure among large numbers of variables is an important topic in many different application areas, such as ecology, neuroscience, genetics. In a graphical model, the dependence structure of a random vector $\boldsymbol{Y} = (Y_1, \dots, Y_p)$ can be represented by a graph G with nodes $\{1, \dots, p\}$, where each node i corresponds to a random variable Y_i and edges represent the probabilistic relationships between nodes.

If there is not an edge connecting nodes i and j it means that, conditionally on $\boldsymbol{Y} \setminus \{Y_i, Y_j\}$, Y_i and Y_j are independent. When $\boldsymbol{Y}$ is assumed to be a Gaussian random vector, the objective of the inference is the precision matrix Σ^{-1}, the inverse of the covariance matrix, which encodes conditional (in)dependencies: $\boldsymbol{\Sigma}^{-1}_{ij} = 0$ if and only if Y_i and Y_j are independent given $\boldsymbol{Y} \setminus \{Y_i, Y_j\}$. For a recent review, see Maathuis et al. (2018).

In the Bayesian setting, a prior distribution is assumed on Σ^{-1} that encourages its off-diagonal entries to be zero or close to zero. Two strategies are commonly employed: shrinkage priors and graph-based priors. The former approach can be understood as a generalization of commonly used shrinkage priors (such as the Lasso prior) in linear regression to positive definite matrices. See, for instance, Wang (2012); Li et al. (2019) and references therein. In the latter approach, instead, a prior is assumed for G and, conditionally to G a prior on Σ^{-1} is assumed such that an absence of the edge between nodes i and j in G implies $\Sigma^{-1}_{ij} = 0$. See Mohammadi and Wit (2015) and references therein. Each approach has its pros and cons. Generally speaking, posterior inference in graph-based models is less efficient because they require transdimensional Markov chain Monte Carlo (MCMC) sampling strategies (Green, 1995) in a huge dimensional parameter space. On the other hand, models based on shrinkage priors usually lead to simpler and more efficient MCMC algorithms, but the estimates of G obtained from Σ^{-1} might be worse (Mohammadi and Wit, 2015).

We propose a novel shrinkage prior for Bayesian graphical modeling, called Fiedler prior, which is particularly useful for estimating sparse precision matrices Σ^{-1} with a block-diagonal structure. We borrow ideas from spectral clustering (Von Luxburg, 2007) and define the prior based on the spectrum of a transformation of the precision matrix. This allows Fiedler prior to enforce block-diagonal structure on the precision matrix.

There exist several methods for sparse covariance matrix estimation based on approximating the precision matrix in a block-diagonal way. These approaches usually follow a two-step procedure, first detecting the block-diagonal structure and then applying the Graphical Lasso (hereafter referred to as G-Lasso) algorithm to each block for estimating the precision matrix (see eg Devijver and Gallopin, 2018).

2 The Bayesian graphical model

Before presenting the main contribution of this work, let us introduce some preliminary definitions and results.

2.1 Graph Laplacian and Fiedler value

Given a weighted graph with weights $W = \{w_{ij}\}_{i,j=1}^{p}$, $w_{ij} \geq 0$, define its unnormalized Laplacian as $L = D - W$, where $D = \text{diag}(\sum_j w_{1j}, \ldots, \sum_j w_{pj})$. The analysis of the eigenvalues $\lambda_1 \leq \ldots \leq \lambda_p$ of L and the associated eigenvectors is formalized in the field of *spectral graph theory*, cf. Spielman (2012).

It is well known that $\lambda_1 = 0$ for any L. The multiplicity of the eigenvalue 0 corresponds to the number of connected components in the graph (see, e.g., Von Luxburg, 2007, Proposition 2). In particular, the graph is connected if and only if the second smallest eigenvalue of L, known as the Fiedler value or algebraic connectivity, satisfies $\lambda_2 > 0$. Moreover, the eigenspace associated with 0 is spanned by the indicator vectors of those components. This is the key motivation underlying spectral clustering.

2.2 The Fiedler prior

In this section, we formalize the Fiedler prior, a prior over partial correlation matrices. The entries of the partial correlation matrix Ω are elements of $[-1, 1]$ which are expressed as a function of Σ^{-1} by

$$\omega_{ij} = -\Sigma^{-1}_{ij} \big/ \sqrt{\Sigma^{-1}_{ii}\Sigma^{-1}_{jj}}. \tag{1}$$

Formally, let $L(|\Omega|)$ be the Laplacian matrix associated to the matrix $|\Omega|$ with entries $|\omega_{ij}|$, and let $\lambda_1(\Omega), \ldots, \lambda_p(\Omega)$ denote its eigenvalues. Then Ω follows the Fiedler prior with parameters $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_p)$ if it has density

$$p(\Omega|\boldsymbol{\delta}) = \frac{1}{Z} \exp\big(-\sum_{j=1}^{p} \delta_j \lambda_j(\Omega)\big) \tag{2}$$

with respect to the standard Lebesgue measure on the space of $[-1, 1]$-valued symmetric matrices. Note that $Z = Z(\boldsymbol{\delta})$ is finite almost surely because the support of $p(\Omega)$ is bounded. Since $\lambda_1 = 0$ for any Ω, we will always set $\delta_1 = 0$. The original idea that initiated the definition and study of this prior is the use of the Fiedler value for penalized maximum likelihood estimation in neural networks (Tam and Dunson, 2020).

To transpose (2) to precision matrices, we use an approach similar to the one in Barnard et al. (2000), who instead work on the covariance matrix Σ. We decompose Σ^{-1} into a partial correlation matrix and an inverse-scale matrix: $\Sigma^{-1} = T\Omega T$, where $T = \mathrm{diag}(\tau_1, \ldots, \tau_p)$. The conditional dependencies can be read equivalently from Ω or Σ^{-1}. Our prior specification is completed by assuming

$$\tau_j \overset{\text{iid}}{\sim} \mathrm{Exp}(\eta), \quad j = 1, \ldots, p, \tag{3}$$

where $\mathrm{Exp}(\eta)$ is the exponential distribution with mean η^{-1}.

As a simple illustration, we compare the marginal distribution on the off-diagonal entries ω_{ij} under G-Lasso and Fiedler priors. Since both priors involve intractable normalizing constants, we use an MCMC algorithm to sample from them. In particular, for the G-Lasso prior we simulate from the prior on Σ^{-1} defined in Wang (2012) and compute ω_{ij} as in (1). For both priors we assume $p = 15$, for the Fiedler prior we fix $\boldsymbol{\delta} = (0, \delta, \delta, \delta, 0, \ldots, 0)$ with $\delta = 25$. For the G-Lasso prior, we employ a double exponential kernel with parameter λ. Figure 1 shows the marginal distributions for different

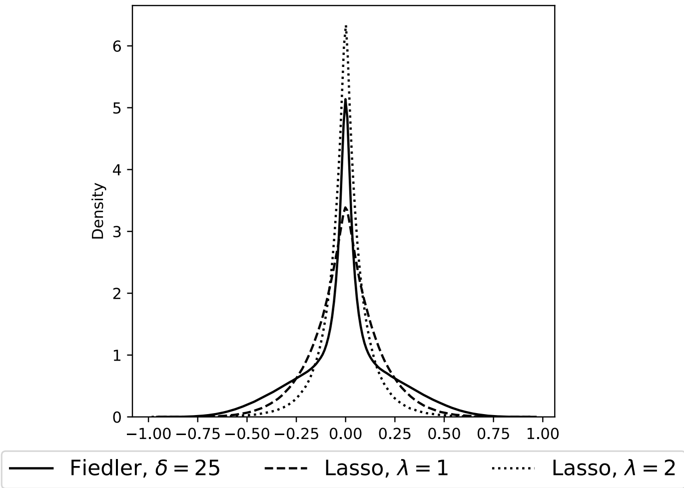

Figure 1: Marginal priors for the off-diagonal entries $\{\omega_{ij}, j > i\}$ under the Fiedler prior and the G-Lasso prior, for different values of the hyperparameters.

values of the parameters. Note that the G-Lasso prior shows the usual tradeoff between local and global shrinkage: to obtain shrinkage for values that are close to 0, also the values that are far from 0 are significantly shrunk (see the tails for $\lambda = 2$). On the contrary, observe how the tails of the Fiedler prior are significantly heavier than the ones of the G-Lasso for both choices of λ, showing that good shrinkage of small values can be achieved without overshrinking the signal of large values.

3 Numerical illustrations

We present a simple simulation study to show the difference between the Fiedler prior and the G-Lasso. We simulated $n = 250$ observations independently from a six-dimensional zero centered normal distribution with precision matrix equal to

$$\Sigma^{-1} = \begin{bmatrix} A, & \mathbf{0} \\ \mathbf{0}, & A \end{bmatrix}, \qquad A = \begin{bmatrix} 3, & 1.5, & 1.5 \\ 1.5, & 3, & 1.5 \\ 1.5, & 1.5, & 3 \end{bmatrix}. \tag{4}$$

Such a model separates the variables into two blocks: the first three and the last three.

We considered different prior specifications for Σ^{-1}. A "well-specified" and a "misspecified" Fiedler prior with respective parameters $\boldsymbol{\delta} = (0, \delta, 0, \dots, 0)$ and $\boldsymbol{\delta} = (0, \delta, \delta, \dots, 0)$. Such parameters $\boldsymbol{\delta}$ imply that they make use of only λ_2, or both λ_2 and λ_3, respectively, which should yield two or three separate groups of variables, respectively. Finally, we considered the G-Lasso with double exponential parameter λ.

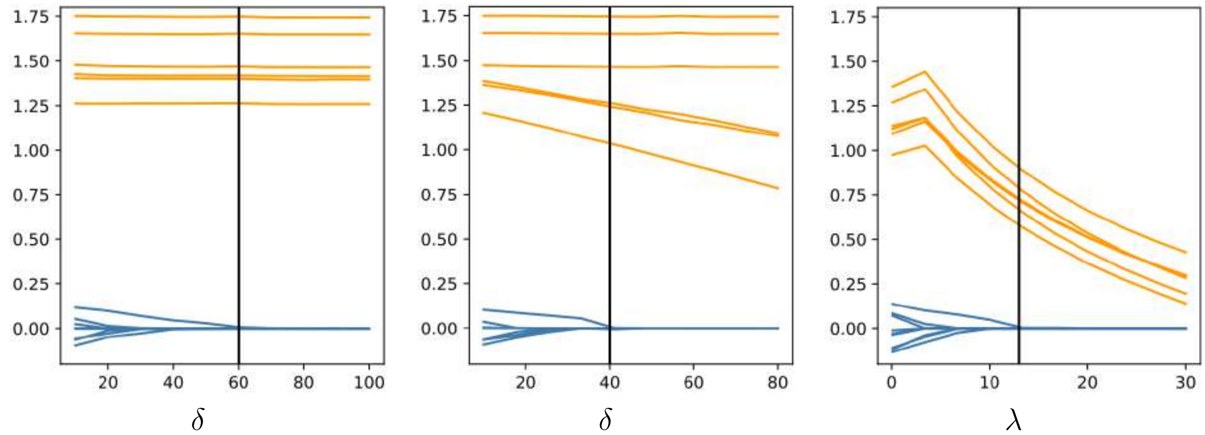

Figure 2: Coefficients of the MAP estimates of Σ^{-1} as a function of the values δ or λ under the tree models: from left to right, well-specified Fiedler, misspecified Fiedler, and G-Lasso. Orange lines and blue lines refer to the estimates of the nonzero and zero off-diagonal elements of Σ^{-1}, respectively. The vertical black line indicates when all the estimates of the zero entries in (4) are below 10^{-5} in absolute value.

We computed the maximum a posteriori (MAP) estimate of Σ^{-1} for various values of δ and λ and looked at the values of the entries of Σ^{-1} as a function of δ and λ.

Figure 2 reports the plots of the "paths" for the tree priors employed. The G-Lasso prior shows the usual overshrinking phenomenon: to estimate values close to zero for the zeros in Σ^{-1}, all the values are shrunk to small values. The well-specified Fiedler behaves correctly: it shrinks to zero the correct terms in Σ^{-1} without "penalizing" the nonzero entries. This shows exactly how the Fiedler prior works: it encourages sparsity only to separate components of the graph associated with Σ^{-1}. Once the components are separated, the other variables are free to assume any large value. Finally, the misspecified Fiedler shows an in-between behavior. In order to recover the two-block structure in Σ^{-1} also some of its nonzero entries are shrunk. This suggests that great care must be taken in carefully choosing the parameter $\boldsymbol{\delta}$.

4 Discussion and future work

In this work, we presented a novel prior for partial correlation matrices, namely the Fiedler prior, and showed an application to Bayesian graphical modeling for Gaussian variables. The Fiedler prior is particularly suited to detect block-diagonal structures.

Several interesting questions are still open. First of all, the choice of parameter $\boldsymbol{\delta}$ seems to be crucial. Assuming a prior distribution on it is unpractical due to the intractable normalizing constant in (2). Hence, a suitable prior elicitation strategy, as well as sensitivity analysis, must be devised. Second, MCMC computation based on gradient information is burdensome due to the need of computing the gradient of the eigendecomposition of Ω. To this end, we might exploit an approximate gradient formulation based on the Rayleigh quotient characterization, as done in Tam and Dunson (2020).

References

Barnard, J., McCulloch, R., and Meng, X.-L. (2000). "Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage." *Statistica Sinica*, 1281–1311.

Devijver, E. and Gallopin, M. (2018). "Block-diagonal covariance selection for high-dimensional Gaussian graphical models." *Journal of the American Statistical Association*, 113(521), 306–314.

Green, P. J. (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination." *Biometrika*, 82(4), 711–732.

Li, Y., Craig, B. A., and Bhadra, A. (2019). "The graphical horseshoe estimator for inverse covariance matrices." *Journal of Computational and Graphical Statistics*, 28(3), 747–757.

Maathuis, M., Drton, M., Lauritzen, S., and Wainwright, M. (2018). *Handbook of graphical models*. CRC Press.

Mohammadi, A. and Wit, E. C. (2015). "Bayesian structure learning in sparse Gaussian graphical models." *Bayesian Analysis*, 10(1), 109–138.

Spielman, D. (2012). "Spectral graph theory." *Combinatorial scientific computing*, (18).

Tam, E. and Dunson, D. (2020). "Fiedler Regularization: Learning Neural Networks with Graph Sparsity." In *Proceedings of ICML 37*, volume 119, 9346–9355. PMLR.

Von Luxburg, U. (2007). "A tutorial on spectral clustering." *Statistics and Computing*, 17(4), 395–416.

Wang, H. (2012). "Bayesian graphical lasso models and efficient posterior computation." *Bayesian Analysis*, 7(4), 867–886.