Notes on Gaussian Graphical Models

Bayesian Statistics

Alessandro Colombi

August 23, 2022

1 Introduction

Disclaimer: These notes are takes from a preliminary draft paper for the work about block graphs (with fixed blocks).

[———— **Introduzione Graphical Models** ————]

Probabilistic graphical modeling is a possible approach to the task of studying the dependence structure among a set of variables. It relies on the concept of conditional independence between variables that is described through a map between a graph and a family of multivariate probability models. When such a family of probabilities is chosen to be Gaussian, those models are known as Gaussian graphical models (Lauritzen 1996). This is the choice made throughout the paper, which is the most common in the literature.

Let $\boldsymbol{X}$ be a p-random vector distributed as a multivariate normal distribution with zero mean and precision matrix $\boldsymbol{K}$, i.e., $N_p\left(\boldsymbol{0}, \boldsymbol{K}^{-1}\right)$; without loss of generality, we assume here $\boldsymbol{X}$ to be centered.

Let $G = (V, E)$ be an undirected graph, where $V = \{1, \dots, p\}$ is the set of nodes and $E \subset \mathscr{E} = \{ (i, j) \mid i < j,\ i, j \in V \}$ is the set of undirected edges. $\boldsymbol{X}$ is said to be Markov with respect to G if, for any edge (i, j) that does not belong to E, the i-th and j-th variables of $\boldsymbol{X}$ are conditionally independent given all the others ($X_i \perp\!\!\!\perp X_j \mid \boldsymbol{X}_{-(ij)}$, where $\boldsymbol{X}_{-(ij)}$ is the random vector containing all elements in $\boldsymbol{X}$ except the i-th and the j-th). Under the normality assumption, the conditional independence relationship between variables can be represented in terms of the null elements of the precision matrix $\boldsymbol{K}$. Specifically, the following equivalence provides an interpretation of the graph

$$X_i \perp\!\!\!\perp X_j \mid \boldsymbol{X}_{-(ij)} \iff (i, j) \notin E \iff k_{ij} = 0. \tag{1}$$

Let $\boldsymbol{Y} = \{\boldsymbol{y}_1, \dots, \boldsymbol{y}_n\}$ be an independent and identically distributed (iid) sample of size n from $\mathscr{N}_p\left(\boldsymbol{0}, \boldsymbol{K}^{-1}\right)$. Usually the underlying graph G is unknown and it is the goal of the statistical inference, along with $\boldsymbol{K}$. Such a process is also known as structural learning. In a Bayesian framework, we set a G-Wishart prior distribution for the precision matrix $\boldsymbol{K}$ (Roverato 2002; Atay-Kayis and Massam 2005) which is attractive as it is conjugate to the likelihood. Following Mohammadi and Wit (2015), we use a *Shape-Inverse Scale* parametrization for such a distribution, that is, we say that $\boldsymbol{K} \sim$ G-Wishart(b, D) if its density is given by

$$P\left(\boldsymbol{K} \mid G, b, D\right) = I_G\left(b, D\right)^{-1} |\boldsymbol{K}|^{\frac{b-2}{2}} \exp\left\{-\frac{1}{2} tr\left(\boldsymbol{K}D\right)\right\} \mathbb{1}_{\mathbb{P}_G}, \tag{2}$$

where $b > 2$ is the Shape parameter, the Inverse Scale matrix D is symmetric and positive definite and $\mathbb{P}_G$ is the space of all $p \times p$ symmetric and positive definite matrices whose null elements are associated to links absent in graph G, i.e., that are Markov with respect to G. Finally, $I_G(b, D)$ is the normalizing constant. In this work, b and D are fixed hyperparameters. Thanks to conjugacy, we have $\boldsymbol{K} \mid G, \boldsymbol{Y} \sim$ G-Wishart $(b + n, D + U)$, where $U = \boldsymbol{Y}^T\boldsymbol{Y}$.

[———— **Recap Priors** ————]

Since the graph G is considered to be a random variable having values in the space $\mathcal{G}$ of all possible undirected graphs with p nodes, we need to specify a prior $\pi(G)$ on it. A common practice is to choose an uniform distribution over $\mathcal{G}$. This is appealing for its simplicity but Jones et al. (2005) noted that it is not a convenient choice to impose sparsity, as it assigns most of its mass to graphs with a "medium" number of edges. Indeed, $\pi_U(G) = 1/|\mathcal{G}| \propto 1/\binom{|\mathcal{E}|}{|E|}$, for each $G \in \mathcal{G}$, where $|\mathcal{G}| = 2^{\binom{p}{2}}$, $|\mathcal{E}| = \binom{p}{2}$. Such a distribution peaks when the number of edges $|E| = p(p-1)/4$ (Scott and Carvalho 2008). [**Questa è un'infromazione un po' troppo specifica. L'ho riportata per averla scritta da qualche parte.**] As a consequence, it is not as non-informative as one usually expect. As an alternative, Mohammadi and Wit (2015) proposed a truncated Poisson distribution on the graph size, $\pi_{tp}(G) \propto \gamma^{|E|}/|E|!$ for each $G \in \mathcal{G}$.

On the other hand, it is known that an undirected graph is uniquely identified by its set of edges $\mathcal{E}$. Therefore it is simpler to define a prior on $\mathcal{E}$, which then naturally induces a prior over $\mathcal{G}$. In this setting, the most natural choice is to assign independent Bernoulli(θ_e) priors to each link $e \in \mathcal{E}$. The Bernoulli parameters θ_e could be different from edge to edge, but one usually assigns a common value, $\theta_e = \theta$, $\forall e \in \mathcal{E}$. For example, Jones et al. (2005) suggest the choice $\theta = 2/(p-1)$ to induce more sparsity in the graph. Scott and Carvalho (2008) place instead a Beta hyperprior on θ, a solution known as multiplicity correction prior. Similarly, Scutari (2013) described a multivariate Bernoulli distribution where edges are not necessarily independent.

[———— **Sampling Strategy** ————]

Bayesian posterior inference of the graph is usually performed through Markov chain Monte Carlo (MCMC) under the conjugate G-Wishart prior distribution on the precision matrix. However, the posterior computation is expensive for general nondecomposable graphs for two main reasons.

Firstly, note that the cardinality of the model space is $|\mathcal{G}| = 2^{\binom{p}{2}}$. Hence, it is large even if a moderate number of variables p is included. In practice, it can not be explicitly numerated but one needs to rely on search algorithm to explore it and learn what links should be included or not. Secondly, the difficulty in the development of efficient methods for structural learning is due to the presence of the G-Wishart prior distribution, which is defined conditionally on a graph G and it is known only up to the intractable normalizing constant $I_G(b, D)$. Actually, an analytic form does exist (Uhler et al. 2018), but the expression is mathematically complex and its implementation is not feasible at the moment (Mohammadi et al. 2021). Explicit formulas do exists for special cases such as complete or decomposable graphs, that are hard to justify from an applied side and increasingly restrictive as the number of nodes increases. In practice, the normalizing constant is usually evaluated by means of numerical approximations such as Monte Carlo approximation (Atay-Kayis and Massam 2005), importance sampler (Roverato 2002; Dellaportas et al. 2003) and the Laplace approximation (Moghaddam et al. 2009; Lenkoski and Dobra 2011). Unfortunately, these methods become unstable with an increasing number of nodes (Jones et al. 2005; Wang and Li 2012). Recent solutions have been proposed in the literature. For example Wang et al. (2015) leverages on the partial analytical structure of the G-Wishart distribution while Mohammadi and Wit (2015) rely on an approximation of the ratio of two normalizing constants when comparing two models. (van den Boom et al. 2022) $I_{G^{-e}}(b, D)/I_G(b, D)$. However, these approaches are suited for comparing models whose graphs differ by a single edge and so they are inappropriate to address block structural learning of our setting.

2 Sampling Strategy for Gaussian Graphical Models (Extended)

The development of efficient methods for structural learning in Gaussian graphical models under the G-Wishart prior has been a wide research field. The main challenge to be faced is that the joint posterior distribution of graph and precision matrix is doubly-intractable (Murray et al. 2006) because it depends on the normalization constant $I_G(b, D)$, defined in (2), which is function of the parameters of interest.

In principle, the precision matrix $\boldsymbol{K}$ can be considered to be a nuisance parameter and integrated

out of the model thanks to conjugacy. Within this framework, both MCMC methods (see Giudici and Castelo (2003) and Bhadra and Mallick (2013)) and stochastic search algorithm (see Jones et al. (2005); Scott and Carvalho (2008); Lenkoski and Dobra (2011)) has been developed. Eliminating $\boldsymbol{K}$ poses non negligible computational issues due to the arise of the G-Wishart posterior normalizing constant $I_G(b+n, D+U)$, which leads to even greater instabilities with respect to the prior normalizing constant $I_G(b, D)$ (Lenkoski and Dobra 2011).
One way to avoid dealing with such an unstable approximation is to set up a chain on the joint space of graph and precision matrix. The latter introduces the need of a Reversible Jump approach but does not eliminate the presence of $I_G(b, D)$ (Giudici and Green 1999; Dobra et al. 2011).

Wrapping up, the main difficulties to carry out inference are the cardinality of the space of all possible graphs $\mathscr{G}$ and the presence of the G-Wishart normalizing constant.

As already mentioned, the dimension of $\mathscr{G}$ grows at combinatorial speed. Even when the number of nodes is limited number, it may be extremely difficult to identify high posterior probability regions. We face the problem by proposing a model space reduction. We develop a chain that visits only the subspace of block structured graphs $\mathscr{B}$, which is in general much smaller. Moreover, existing methods are usually able to modify only one link at each step of the chain. Nevertheless, to guarantee a block structure compatible with our hypotheses, links can not be modified at will, at least, not in the space $\mathscr{B}$. Our approach consists of mapping through ρ^{-1} the current graph in its multigraph representation G_B where block of links are represented by a single edge. There we can use standard tools for structural learning and then we can map the new graph, G'_B, back into its block form G'. The last step requires some care due to the constraints imposed on the precision matrix through the G-Wishart distribution.

The G-Wishart normalizing constant $I_G(b, D)$ plays a crucial role in the computation of the Bayes factors in model comparison. Although an analytic form does exist (Uhler et al. 2018), the expression is mathematically complex and its implementation is not feasible at the moment. Explicit formulas do exists for special cases such as complete or decomposable graphs, but in general the only way to evaluate $I_G(b, D)$ is by means of numerical approximations. Roverato (2002) and Dellaportas et al. (2003) developed an importance sampler, Lenkoski and Dobra (2011) and Moghaddam et al. (2009) instead addressed the problem by proposing a Laplace transform approximation. However, the reference method is the Monte Carlo approximation developed by Atay-Kayis and Massam (2005). Although it is a valid procedure, modern methods prefer to avoid this calculation which as been proved to be unstable in high dimensional problems and it may require an exaggerate amount of iterations to reduce the Monte Carlo variance which makes it infeasible due to the computational burden it requires. See Jones et al. (2005) and Wang and Li (2012) for further details.
A reasonable question is if there is a simple way to at least approximate $I_G(b, D)/I_{G'}(b, D)$, which is, in practice, the calculation one wants to compute. A possible estimate was first verified numerically (Mohammadi and Wit 2012) and then proved in Mohammadi et al. (2021). Let $G = (V, E)$ and set D equal to $\mathbb{I}_p$. Than define a new graph $G' = (V, E')$ with $E' = E \backslash e$, which is the graph obtained by removing edge $e = (i, j)$ form G. In this case, the normalizing constant ratio can be approximated as

$$\frac{I_{G^{-e}}(b, \mathbb{I}_p)}{I_G(b, \mathbb{I}_p)} \approx \frac{1}{2\sqrt{\pi}} \frac{\Gamma\left(\frac{b+d}{2}\right)}{\Gamma\left(\frac{b+d+1}{2}\right)} \tag{3}$$

where d is the number of paths of length two linking the two end points i and j of e. What we want to stress is that this estimate is valid as far as the two compared graphs differ only for a single link. What we want to achieve is, instead, a MCMC method that modifies more that one link at a time. In general, we would like to change an arbitrary number of edges. In this case, we are not aware of any estimate for the G-Wishart normalizing constants ratio. Wrapping up, the joint posterior distribution of graph and precision matrix is doubly-intractable and none of the existing method is reliable enough in evaluating $I_G(d, D)$.

Within the literature about the problem of dealing with doubly-intractable distributions, we focus on the researches carried out first by Müller et al. (2007) and then by Murray et al. (2006). The

method proposed by the authors is to introduce an auxiliary variable that cancels out the intractable constant when computing the acceptance probability. The algorithm is asymptotically exact as far as a direct sample to draw from the target distribution is available. See Park and Haran (2020) for an exhaustive review of other approaches.

The problem of incorporating the Exchange algorithm by Murray et al. (2006) within a sampling strategy to draw samples from a Gaussian graphical model is not straightforward. Indeed, the presence of the G-Wishart distribution requires a trans-dimensional extension of the Exchange algorithm. Wang and Li (2012) and Lenkoski (2013) proposed two different MCMC method to couple with this problem. The method by Wang and Li (2012) is on based on the partial analytic structure of the G-Wishart and it is very appealing since it does not require any proposal tuning nor any matrix completion operation (Atay-Kayis and Massam 2005, prop. 2). However, it strongly relies on the possibility of writing down an explicit formula for the full conditional of the elements of $\boldsymbol{K}$. Such results are presented in Roverato (2002) but they can be handled in practice only in special cases. Recent developments include the WWA method by van den Boom et al. (2022) which improves the Double Reversible Jump procedure of Hinne et al. (2014) by leveraging on the delayed acceptance MCMC (Christen and Fox 2005) and an informed proposal (Zanella 2020) distribution on the graph space that enables embarrassingly parallel computation. As a consequence, the WWA reduces the frequency of the expensive sampling from the G-Wishart distribution resulting in fast convergence and good MCMC mixing. Both the MCMC method of Wang and Li (2012) and the WWA are feasible if at each step of the graph only one link of the graph is modified. It is extremely complicated, if possible at all, to generalize them to a general framework where an arbitrary number of links is modified.

References

Atay-Kayis A, Massam H (2005) A Monte Carlo method for computing the marginal likelihood in nondecomposable Gaussian graphical models. Biometrika 92:317–335

Bhadra A, Mallick BK (2013) Joint high-dimensional Bayesian variable and covariance selection with an application to eQTL analysis. Biometrics 69(2):447–457

van den Boom W, Beskos A, Iorio MD (2022) The g-wishart weighted proposal algorithm: Efficient posterior computation for gaussian graphical models. Journal of Computational and Graphical Statistics 0(0):1–10, DOI 10.1080/10618600.2022.2050250, URL https://doi.org/10.1080/10618600.2022.2050250, https://doi.org/10.1080/10618600.2022.2050250

Christen JA, Fox C (2005) Markov chain monte carlo using an approximation. Journal of Computational and Graphical Statistics 14(4):795–810, URL http://www.jstor.org/stable/27594150

Dellaportas P, Giudici P, Roberts G (2003) Bayesian inference for nondecomposable graphical Gaussian models. Sankhyā: The Indian Journal of Statistics 65(1):43–55

Dobra A, Lenkoski A, Rodriguez A (2011) Bayesian inference for general Gaussian graphical models with application to multivariate lattice data. Journal of the American Statistical Association 106(496):1418–1433

Giudici P, Castelo R (2003) Improving Markov chain Monte Carlo model search for data mining. Machine learning 50(1-2):127–158

Giudici P, Green P (1999) Decomposable graphical Gaussian model determination. Biometrika 86(4):785–801

Hinne M, Lenkoski A, Heskes TM, van Gerven M (2014) Efficient sampling of gaussian graphical models using conditional bayes factors. Stat 3:326 – 336

Jones B, Carvalho C, Dobra A, Hans C, Carter C, West M (2005) Experiments in stochastic computation for high-dimensional graphical models. Statistical Science 20:388–400

Lauritzen SL (1996) Graphical models. Oxford University Press, Oxford

Lenkoski A (2013) A direct sampler for G-Wishart variates. Stat 2(1):119–128

Lenkoski A, Dobra A (2011) Computational aspects related to inference in Gaussian graphical models with the G-Wishart prior. Journal of Computational and Graphical Statistics 20(1):140–157

Moghaddam B, Khan E, Murphy KP, Marlin BM (2009) Accelerating bayesian structural inference for non-decomposable gaussian graphical models. Advances in Neural Information Processing Systems 22

Mohammadi A, Wit EC (2012) Gaussian graphical model determination based on birth-death mcmc inference

Mohammadi A, Wit EC (2015) Bayesian structure learning in sparse Gaussian graphical models. Bayesian Analysis 10(1):109–138

Mohammadi R, Massam H, Letac G (2021) Accelerating bayesian structure learning in sparse gaussian graphical models. J Amer Statist Assoc 0(0):1–14

Müller P, Parmigiani G, Rice K (2007) FDR and Bayesian multiple comparisons rules. In: Bernardo JM, Bayarri M, Berger J, Dawid A, Heckerman D, Smith A, West M (eds) Bayesian Statistics 8, Oxford University Press, Oxford

Murray I, Ghahramani Z, MacKay D (2006) Mcmc for doubly-intractable distributions. In: Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06), AUAI Press, pp 359–366

Park J, Haran M (2020) A function emulation approach for doubly intractable distributions. Journal of Computational and Graphical Statistics 29(1):66–77, DOI 10.1080/10618600.2019.1629941, URL https://doi.org/10.1080/10618600.2019.1629941, https://doi.org/10.1080/10618600.2019.1629941

Roverato A (2002) Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. Scandinavian Journal of Statistics 29(3):391 – 411

Scott J, Carvalho C (2008) Feature-inclusion stochastic search for Gaussian graphical models. Journal of Computational and Graphical Statistics 17(4):790–808

Scutari M (2013) On the prior and posterior distributions used in graphical modelling. Bayesian Analysis 8(3):505–532

Uhler C, Lenkoski A, Richards D (2018) Exact formulas for the normalizing constants of wishart distributions for graphical models. The Annals of Statistics 46(1):90–118

Wang H, Li SZ (2012) Efficient Gaussian graphical model determination under G-Wishart prior distributions. Electronic Journal of Statistics 6:168–198

Wang H, et al. (2015) Scaling it up: Stochastic search structure learning in graphical models. Bayesian Analysis 10(2):351–377

Zanella G (2020) Informed proposals for local mcmc in discrete spaces. Journal of the American Statistical Association 115(530):852–865, DOI 10.1080/01621459.2019.1585255, URL https://doi.org/10.1080/01621459.2019.1585255, https://doi.org/10.1080/01621459.2019.1585255