Goodness of fit of logistic models for random graphs

Pierre Latouche, Stéphane Robin, Sarah Ouadah

Goodness of fit of logistic models for random graphs

Pierre Latouche[1], Stéphane Robin[2,3], Sarah Ouadah[2,3]

([1]) Laboratoire SAMM, EA 4543, Université Paris 1 Panthéon-Sorbonne, France
([2]) AgroParisTech, UMR 518, MIA, Paris, France
([3]) INRA, UMR 518, MIA, Paris, France

August 4, 2015

Abstract

Logistic models for random graphs are commonly used to study binary networks when covariate information is available. After estimating the logistic parameters, one of the main questions which arises in practice is to assess the goodness of fit of the corresponding model. To address this problem, we add a general term, related to the graphon function of W-graph models, to the logistic function. Such an extra term aims at characterizing the residual structure of the network, that is not explained by the covariates. We approximate this new generic logistic model using a class of models with blockwise constant residual structure. This framework allows to derive a Bayesian procedure from a model based selection context using goodness-of-fit criteria. All these criteria depend on marginal likelihood terms for which we do provide estimates relying on two series of variational approximations. Experiments on toy data are carried out to assess the inference procedure. Finally, two real networks from social sciences and ecology are studied to illustrate the proposed methodology.

keywords : Random graphs; logistic regression; W-graph model; variational approximations

1 Introduction

Networks are now used in many scientific fields [14, 16, 30, 32, 37, 39] from biology [2, 4, 22, 29] to historical sciences [19, 34] and geography [12]. Indeed, while being simple data structures, they are yet capable of describing complex interactions between entities of a system. A lot of effort has been put, especially in social sciences, in developing methods to characterize the heterogeneity of these networks using latent variables and/or covariate information. Latent variable techniques [13, 28] usually associate a hidden variable to each node of a network such that the construction of edges involve mixture models. Because

nodes can have different latent variables, they can have various connectivity patterns within the network. A long series of papers has focused in the last ten years on the stochastic block model (SBM) [30, 36]. The model assumes that nodes are spread in unknown clusters and that the probability of connection of two nodes depends exclusively on the clusters they belong to. Note that extensions have been proposed for SBM to deal for instance with valued edges [27]. This alternative approach allows the use of covariates to explain the presence of interactions between nodes. The latent position model (LPM) [16] and the latent position cluster model (LPCM) [14] also allow to consider both latent variables and covariates [27, 40]. Their goal is to integrate the two sources of heterogeneity in a principle manner and assumptions are usually made regarding the hidden connectivity patterns. In LPCM for instance, latent variables allow to model communities only where two nodes of the same community are more likely to be connected.

In this paper, we tackle a different problem. Thus, we consider standard logistic models which are highly used in practice to deal with covariates in networks, assuming edges to be independent. Our framework aims at allowing practitioners to assess the goodness of fit of their estimated models, *i.e.* testing the presence of heterogeneity in the network not accounted for by the fitted model. It relies on two of the most flexible random graph models with latent variables, namely the SBM and the *W-graph model*, to characterize the residual structure not explained by the covariates.

Usual random graph models for binary networks, like SBM, can be seen as special cases of the W-graph model. This model is characterized by a function W called *graphon* where $W(u, v)$ is the probability for two nodes, with latent coordinates u and v, sampled from an uniform distribution over $[0, 1]$, to connect. As shown in [26], it is the limiting adjacency matrix of the network. This result comes from graph limit theory for which Diaconis and Janson [11] gave a proper definition using Aldous-Hoover theorem, which is an extension of deFinetti's theorem to exchangeable arrays. Until recently, few inference techniques had been proposed to infer the graphon function of a network. The earliest reference is Kallenberg [20]. Since then, both parametric [15, 31] and non parametric [9] techniques have been developed. Graphon inference is a particularly challenging problem which has received strong attention is the last few years [1, 3, 9, 38]. In particular, we point out the work of Latouche and Robin [25] who used a VBEM procedure to approximate the graphon function as an average of SBM models with increasing number of blocks.

In this paper, we add a general term, related to the graphon function of W-graph models, to the logistic regression model. This generic term allows to encode the residual structure present in the data, not explained by the covariates. Unfortunately, exact inference of this new logistic model for networks is not tractable and therefore we propose to rather consider a series of models with blockwise constant residual structures. Within this framework, after introducing prior distributions, the fit can be evaluated from a Bayesian model averaging context and goodness-of-fit criteria are introduced. All these criteria depend on marginal likelihood terms which are not tractable. To tackle this issue, two series of variational approximations are considered and estimates are derived.

In Section 2, we take a general point of view and the discussion is to the point to help introducing the Bayesian testing procedure. Technical issues and theoretical aspects are addressed in Section 3. Finally, toy and real data sets are analyzed in Section 4 and 5 respectively to illustrate the proposed methodology.

2 Assessing goodness-of-fit

We consider a set of n individuals among which interactions are observed. The observed interaction network is encoded in the binary adjacency matrix $Y = (Y_{ij})_{1\leq i,j\leq n}$ where Y_{ij} is 1 if nodes i and j are connected, and 0 otherwise. We further assume that a d-dimensional vector, $d \geq 1$, of covariates x_{ij} is available for each pair of nodes. In the following, we denote as $X = (x_{ij})_{1\leq i,j\leq n}$ the set of all covariates.

2.1 Logistic regression and residual structure

The influence of the covariates on the network topology can be easily accounted for using a logistic regression model. Such a model assumes that the edges (Y_{ij}) are independent with respective distribution

$$H_0 : \qquad Y_{ij} \sim \mathcal{B}\left[g(x_{ij}^\intercal \beta + \alpha)\right],$$

where $\beta \in \mathbb{R}^d$, $\alpha \in \mathbb{R}$, g stands for the logistic function $g(t) = 1/(1+\exp(-t))$, $t \in \mathbb{R}$. Our goal is to assess the goodness of fit of Model H_0. Note that the network structure does note explicitly appear in this model, as edges are considered as independent outcomes of a (generalized) linear model.

To assess the fit of Model H_0, we define a generic alternative network model. The alternative we consider is inspired from the graphon model. More precisely, we consider the model

$$H_1 : \qquad Y_{ij} \sim \mathcal{B}\left[g(x_{ij}^\intercal \beta + \phi(U_i, U_j))\right],$$

where the $(U_i)_{1\leq i\leq n}$ are independent unobserved latent variables, with uniform distribution over the $(0, 1)$ interval. The non-constant function $\phi : (0, 1)^2 \mapsto \mathbb{R}$ encodes a residual structure in the network, that is not accounted for by Model H_0. Note that, in absence of covariate, this model corresponds to a W-graph ([26]) with graphon function $g \circ \phi$. Model H_0 corresponds to the case where the residual function ϕ is constant.

The inference of the function ϕ in Model H_1 is not an easy task and, following [25] and [1], we consider a class of blockwise constant ϕ function. More precisely, we define the Model

$$M_K : \qquad Y_{ij} \sim \mathcal{B}\left[g(x_{ij}^\intercal \beta + Z_i^\intercal \alpha Z_j)\right], \tag{1}$$

where α is a $K \times K$ real matrix ($K \geq 1$) and where the $(Z_i)_{1\leq i\leq n}$ are independent vectors with K coordinates, all zero except one. We denote π_k ($1 \leq k \leq K$) the probability that the kth coordinate is non-zero. Briefly speaking, each vector Z_i has multinomial

distribution $\mathcal{M}(1, \pi)$ where $\pi = (\pi_k)_{1 \leq k \leq K}$. The set of parameters of such a model is $\theta = (\beta, \pi, \alpha)$. Note that in the absence of covariate, this model corresponds exactly to a SBM model.

Model H_0 is then equivalent to Model M_1 so the goodness-of-fit problem can be rephrased as the comparison between Model H_0 and H_1', where

$$H_0 = M_1 \qquad \text{and} \qquad H_1' = \bigcup_{K \geq 2} M_K.$$

2.2 Bayesian model comparison

Now, we are given a series of Models M_K ($K \geq 1$) indexed by K which characterize H_0 and H_1'. In this paper, we propose to compare H_0 and H_1' using a Bayesian model comparison framework.

Thus, each Model M_K is associated to a prior probability $p(M_K)$. The parameter θ is then drawn conditionally on M_K according to the prior distribution $p(\theta|M_K)$. Given θ, M_K and the given set X of covariates, the graph is finally assumed to be sampled according to Model (1). In this framework the prior probability of Models H_0 and H_1' are

$$p(H_0) = p(M_1) \qquad \text{and} \qquad p(H_1') = \sum_{K \geq 2} p(M_K).$$

Moreover, the posterior probability of Model M_K is

$$p(M_K|Y) = \frac{p(Y|M_K)p(M_K)}{p(Y)} = \frac{p(Y|M_K)p(M_K)}{\sum_{K' \geq 1} p(Y|M_{K'})p(M_{K'})}. \tag{2}$$

The goodness of fit of Model H_0 can then be assessed by computing the posterior probability of H_0:

$$p(H_0|Y) = p(M_1|Y). \tag{3}$$

The Bayes factor [21] between Models H_0 and H_1' can be computed in a similar way as

$$B_{01} = \frac{p(Y|H_0)}{p(Y|H_1')} \qquad \text{where} \quad p(Y|H_1') = \frac{1}{p(H_1')} \sum_{K \geq 2} p(M_K)p(Y|M_K). \tag{4}$$

3 Inference

The goodness-of-fit criteria introduced in the previous section all depend on marginal likelihood terms $p(Y|M_K)$ which have to be estimated from the data in practice. This is the object of this section. The prior distributions $p(M_K)$ and $p(\theta|M_K)$ are first introduced. A variational three steps optimization scheme, based on global and local variational methods, is then derived for inference.

In the following, we focus on undirected networks and therefore both the adjacency matrix Y and the matrix X of covariates are symmetric: $Y_{ij} = Y_{ji}$ and $x_{ij} = x_{ji}, \forall i \neq j$. The complete derivation of the model and the inference procedure in the directed case are given as supplementary materials. Moreover, we do not consider self-loops, *i.e.* the connection of a node to itself and therefore the pairs $(i, i), \forall i$ are discarded from the sums and products involved.

3.1 Prior distributions

With no prior information on which model should be preferred, we give equal weights $p(H_0) = p(H_1') = 1/2$ to H_0 and H_1'. Therefore, $p(M_1) = 1/2$. Alternative choices can be made by integrating expert knowledge at hand. Recall that $p(H_1') = \sum_{K \geq 2} p(M_K)$.

For Model M_K, the prior distribution over the model parameters in θ is defined as a product of conjugate prior distributions over the different sets of parameters: $p(\theta|M_K) = p(\beta|M_K)p(\pi|M_K)p(\alpha|M_K)$. Since π is involved in a multinomial distribution to sample the vectors Z_i, a Dirichlet prior distribution is chosen

$$p(\pi|M_K) = \mathrm{Dir}(\pi; e),$$

where e is a vector with K components such that $e_k = e_0 > 0, \forall k \in \{1, \ldots, K\}$. Note that fixing $e_0 = 1/2$ induces a non-informative Jeffreys prior distribution which is known to be proper [18]. It is also possible to obtain a uniform distribution over the $K-1$ dimensional simplex by setting $e_0 = 1$.

In order to characterize the d-dimensional regression vector β, a Gaussian distribution is considered

$$p(\beta|\eta, M_K) = \mathcal{N}(\beta;\ 0, \frac{I_d}{\eta}) = \prod_{j=1}^{d} \mathcal{N}(\beta_j;\ 0, \frac{1}{\eta}),$$

with I_d the $d \times d$ identity matrix and $\eta > 0$ a parameter controlling the inverse variance. Similarly, the matrix α is modeled using a product of Gaussian distributions with $\gamma > 0$ controlling the variance

$$p(\alpha|\gamma, M_K) = \prod_{k \leq l}^{K} \mathcal{N}(\alpha_{kl};\ 0, \frac{1}{\gamma}).$$

Since we focus on undirected networks, α has to be symmetric and therefore the product involves the $k \leq l$ terms of α. In the directed case (see supplementary materials), the product is over all terms k, l and the vec operator, which stacks the columns of a matrix into a vector, is used to simplify the calculations.

Finally, Gamma distributions are considered for γ

$$p(\gamma|M_K) = \mathrm{Gam}(\gamma;\ a_0, b_0), \quad a_0, b_0 > 0,$$

and η

$$p(\eta|M_K) = \mathrm{Gam}(\eta;\ c_0, d_0), \quad c_0, d_0 > 0.$$

By construction, Gamma distributions are informative. In order to limit the influence on the posterior distributions, the hyperparameters controlling the scale (a_0, c_0) and rate (b_0, d_0) are usually set to low values in the literature.

The choice of modeling the prior information on the parameters α and β from such Gaussian-Gamma distributions has been widely considered both in standard Bayesian linear regression and Bayesian logistic regression (see for instance [6, 7]). The prior distributions $p(\beta|M_K)$ and $p(\alpha|M_K)$ are then obtained by marginalizing over $p(\eta|M_K)$ and $p(\gamma|M_K)$ respectively. This results in prior distributions from the class of generalized hyperbolic distributions. For more details, we refer to [8].

In the following, and in order to simplify the notations, the dependency on M_K is omitted in the prior and posterior distributions.

3.2 Variational approximations

Denoting Z the set of all latent vectors (Z_i), the marginal log-likelihood of Model M_K, also called the integrated observed data log-likelihood, is given by

$$\log p(Y|M_K) = \log \left\{ \sum_Z \int p(Y|Z,\alpha,\beta) p(Z|\pi) p(\alpha|\gamma) p(\beta|\eta) p(\pi) p(\gamma) p(\eta) \mathrm{d}\pi \mathrm{d}\alpha \mathrm{d}\beta \mathrm{d}\gamma \mathrm{d}\eta \right\}. \tag{5}$$

It requires a marginalization over the prior distributions of all parameters. In particular, it involves testing all the K^n configurations of Z. Unfortunately, (5) is not tractable and therefore we propose to rely on variational approximations for inference purposes. Let us first consider the global variational decomposition

$$\log p(Y|M_K) = \mathcal{L}_K(q) + \mathrm{KL}\left(q(\cdot)||p(\cdot|Y,M_K)\right). \tag{6}$$

Maximizing the functional $\mathcal{L}_K(\cdot)$, which is a lower bound of $\log p(Y|M_K)$, with respect to the distribution $q(\cdot)$, is equivalent to minimizing the Kullback-Leibler divergence between $q(\cdot)$ and the unknown posterior distribution $p(\cdot|Y)$. $\mathcal{L}_K(\cdot)$ is given by

$$\mathcal{L}_K(q) = \sum_Z \int q(Z,\pi,\alpha,\beta,\gamma,\eta) \log \frac{p(Y,Z,\pi,\alpha,\beta,\gamma,\eta)}{q(Z,\pi,\alpha,\beta,\gamma,\eta)} \mathrm{d}\pi \mathrm{d}\alpha \mathrm{d}\beta \mathrm{d}\gamma \mathrm{d}\eta.$$

In order to maximize the lower bound, we assume that the distribution can be factorized as follows:

$$q(Z,\pi,\alpha,\beta,\gamma,\eta) = q(\pi)q(\alpha)q(\beta)q(\gamma)q(\eta) \prod_{i=1}^{n} q(Z_i).$$

Unfortunately, $\mathcal{L}_K(\cdot)$ is still intractable due to the logistic function in $p(Y|Z,\alpha,\beta)$. Following the work of [17], a tractable lower bound is derived.

Proposition 1 *Given any $n \times n$ positive real matrix $\xi = (\xi_{ij})_{1 \leq i,j \leq n}$, a lower bound of the first lower bound is given by*

$$\log p(Y|M_K) \geq \mathcal{L}_K(q) \geq \mathcal{L}_K(q;\, \xi),$$

where

$$\mathcal{L}_K(q;\,\xi) = \sum_Z \int q(Z,\pi,\alpha,\beta,\gamma,\eta)\log\frac{\sqrt{h(Z,\alpha,\beta,\xi)}p(Z,\pi,\alpha,\beta,\gamma,\eta)}{q(Z,\pi,\alpha,\beta,\gamma,\eta)}\mathrm{d}\pi\mathrm{d}\alpha\mathrm{d}\beta\mathrm{d}\gamma\mathrm{d}\eta,$$

and

$$\log h(Z,\alpha,\beta,\xi) = \sum_{i\neq j}^{n}\Big\{(Y_{ij}-\frac{1}{2})(Z_i^\intercal\alpha Z_j + x_{ij}^\intercal\beta) + \log g(\xi_{ij}) - \frac{\xi_{ij}}{2} \\ - \lambda(\xi_{ij})\left((Z_i^\intercal\alpha Z_j + x_{ij}^\intercal\beta)^2 - \xi_{ij}^2\right)\Big\},$$

with $\xi_{ij}\in\mathbb{R}^+$, $\xi_{ij}=\xi_{ji}$. *Moreover,* $\lambda(\xi_{ij}) = \left(g(\xi_{ij})-1/2\right)/(2\xi_{ij})$, g *being the logistic function.*

The proof is given in Appendix A.1. The quality of the lower bound $\mathcal{L}_K(q;\,\xi)$, which was obtained through a series of Taylor expansions, clearly depends on the choice of the matrix ξ. As we shall see in Section 3.2.2, ξ can be estimated from the data to obtain tight bounds.

3.2.1 Variational Bayes EM

For now, we assume that the matrix ξ is fixed and we rely on $\mathcal{L}_K(q;\,\xi)$ as a lower bound of $\log p(Y|M_K)$. In order to maximize the lower bound, a VBEM algorithm [5] is applied on $\mathcal{L}_K(q;\,\xi)$. This optimization scheme is iterative and is related to the EM algorithm [10]. Keeping all distributions fixed except one, the bound is maximized with respect to the remaining distribution. This procedure is repeated in turn until convergence of the bound. The optimization of the distribution $q(Z)$ over the latent variables usually refers to the variational E step. The updates of $q(\pi)$, $q(\alpha)$, $q(\beta)$, $q(\gamma)$, and $q(\eta)$ refer here to the variational M step. Proposition 2 provides the update formula of the E-step and Propositions 3 to 7 provide these of the M-step. The corresponding proofs are given in Appendix A.2 to A.7.

Proposition 2 *The variational E update step for each distribution* $q(Z_i)$ *is given by:*

$$q(Z_i) = \mathcal{M}(Z_i;\,1,\tau_i),$$

where $\sum_{k=1}^K \tau_{ik} = 1$ *and*

$$\tau_{ik}\propto\exp\Bigg\{\sum_{l=1}^K(m_\alpha)_{kl}\sum_{j\neq i}^n\Big((Y_{ij}-\frac{1}{2}) - 2\lambda(\xi_{ij})x_{ij}^\intercal m_\beta\Big)\tau_{jl} - \sum_{l=1}^K\mathbb{E}_{\alpha_{kl}}[\alpha_{kl}^2]\sum_{j\neq i}^n\lambda(\xi_{ij})\tau_{jl} \\ + \psi(e_k^n) - \psi\Big(\sum_{l=1}^K e_l^n\Big)\Bigg\}.$$

$\psi(\cdot)$ denotes the digamma function which is the logarithmic derivative of the gamma function.

Proposition 3 *The variational M update step for the distribution $q(\pi)$ is given by:*

$$q(\pi) = Dir(\pi;\, e^n),$$

where, $\forall k \in \{1, \ldots, K\}$, $e_k^n = e_0 + \sum_{i=1}^n \tau_{ik}$, τ_{ik} being given by Proposition 2.

Proposition 4 *The variational M update step for the distribution $q(\beta)$ is given by:*

$$q(\beta) = \mathcal{N}(\beta;\, m_\beta, S_\beta),$$

where

$$S_\beta^{-1} = \frac{c_n}{d_n} I_d + \sum_{i \neq j}^{n} \lambda(\xi_{ij}) x_{ij} x_{ij}^\intercal,$$

and

$$m_\beta = S_\beta \frac{1}{2} \sum_{i \neq j}^{n} \left(Y_{ij} - \frac{1}{2} - 2\lambda(\xi_{ij}) \tau_i^\intercal m_\alpha \tau_j \right) x_{ij}.$$

Proposition 5 *The variational M update step for the distribution $q(\gamma)$ is given by:*

$$q(\gamma) = \mathrm{Gam}(\gamma;\, a_n, b_n),$$

where $a_n = a_0 + \frac{K(K+1)}{4}$ and $b_n = b_0 + \frac{1}{2} \sum_{k \leq l}^{K} \mathbb{E}_{\alpha_{kl}}[\alpha_{kl}^2]$.

Proposition 6 *The variational M update step for the distribution $q(\eta)$ is given by:*

$$q(\eta) = \mathrm{Gam}(\eta;\, c_n, d_n),$$

where $c_n = c_0 + \frac{d}{2}$ and $d_n = d_0 + \frac{1}{2}\mathrm{Tr}(S_\beta) + \frac{1}{2} m_\beta^\intercal m_\beta$, S_β and m_β being given by Proposition 4.

Proposition 7 *The variational M update step for the distribution $q(\alpha)$ is given by:*

$$q(\alpha) = \prod_{k \neq l}^{K} \mathcal{N}\left(\alpha_{kl}; (m_\alpha)_{kl}, (\sigma_\alpha^2)_{kl}\right),$$

where

$$(\sigma_\alpha^2)_{kk}^{-1} = \frac{a_n}{b_n} + \sum_{i \neq j}^{n} \lambda(\xi_{ij}) \tau_{ik} \tau_{jk}, \forall k,$$

$$(\sigma_\alpha^2)_{kl}^{-1} = \frac{a_n}{b_n} + 2 \sum_{i \neq j}^{n} \lambda(\xi_{ij}) \tau_{ik} \tau_{jl}, \forall k \neq l,$$

$$(m_\alpha)_{kk} = (\sigma_\alpha^2)_{kk} \sum_{i \neq j}^{n} \left(\frac{1}{2}(Y_{ij} - \frac{1}{2}) - \lambda(\xi_{ij}) x_{ij}^\intercal m_\beta \right) \tau_{ik} \tau_{jk}, \forall k,$$

$$(m_\alpha)_{kl} = (\sigma_\alpha^2)_{kl} \sum_{i \neq j}^{n} \left((Y_{ij} - \frac{1}{2}) - 2\lambda(\xi_{ij}) x_{ij}^\intercal m_\beta \right) \tau_{ik} \tau_{jl}, \forall k \neq l.$$

3.2.2 Optimization of ξ

So far, we have seen how the lower bound $\mathcal{L}_K(q;\,\xi)$ of $\log p(Y|M_K)$ could be maximized with respect to the distribution $q(Z, \pi, \alpha, \beta, \gamma, \eta)$. However, we have not addressed yet how ξ could be estimated from the data. Given a distribution $q(\cdot)$, we propose to maximize $\mathcal{L}_K(q;\,\xi)$ with respect to each variable ξ_{ij} in order to obtain the tightest bound $\mathcal{L}_K(q;\,\xi)$ of $\log p(Y|M_K)$. This follows the work of [7] on Bayesian hierarchical mixture of experts and [23, 24] on the overlapping stochastic block model. As shown in the following proposition, this leads to new estimates $\widehat{\xi}_{ij}$ of ξ_{ij}.

Proposition 8 *The estimate $\widehat{\xi}_{ij}$ of ξ_{ij} maximizing $\mathcal{L}_K(q;\,\xi)$ is given by*

$$\xi_{ij} = \sqrt{\sum_{k,l}^{K} \tau_{ik}\tau_{jl}\mathbb{E}_{\alpha_{kl}}[\alpha_{kl}^2] + 2\sum_{k,l}^{K} \tau_{ik}\tau_{jl}(m_\alpha)_{kl}x_{ij}^\intercal m_\beta + \mathrm{Tr}(x_{ij}x_{ij}^\intercal(S_\beta + m_\beta m_\beta^\intercal))}.$$

Note that $\widehat{\xi}_{ij} = \widehat{\xi}_{ji}, \forall i \neq j$ since the networks considered are undirected.

This gives rise to a three steps optimization scheme. Given a matrix ξ, the variational E and M steps of the VBEM algorithm are used to maximize $\mathcal{L}_K(q;\,\xi)$ with respect to $q(\cdot)$. This distribution is then held fixed and the bound is maximized with respect to ξ. These three steps are repeated until convergence of the lower bound. The proof is given in Appendix A.9.

3.3 Estimation

Goodness-of-fit For any K, we have seen how variational techniques could be used to approximate the marginal log-likelihood $\log p(Y|M_K)$ using a lower bound $\widehat{\mathcal{L}}_K := \max_{q,\xi} \mathcal{L}_K(q, \xi)$. As exposed in Section 2.1, our goodness-of-fit procedure relies on the posterior probability of K, that is $p(M_K|Y)$. Indeed, this posterior distribution cannot be derived in a exact manner but, as shown in [35], the distribution $\widehat{p}(M_K|Y)$ that minimizes the Kullback-Leibler divergence with $p(M_K|Y)$ satisfies

$$\widehat{p}(M_K|Y) \propto p(M_K)\exp\{\widehat{\mathcal{L}}_K\}.$$

The approximate posterior probability of H_0 is then $\widehat{p}(H_0|Y) = \widehat{p}(M_1|Y)$ and the corresponding approximate posterior Bayes factor $\widehat{B}_{01}$, defined in (4), can be computed is the same manner.

The following proposition, which is proved in Appendix A.8, shows that many terms of $\mathcal{L}_K(q;\,\xi)$ vanish, when computed after a specific optimization step, so that the lower bound takes a simpler form.

Proposition 9 *If computed right after the variational M step, the lower bound is given by*

$$\begin{aligned}\mathcal{L}_K(q;\,\xi) = \frac{1}{2}\sum_{i\neq j}^{n}\left\{\log g(\xi_{ij}) - \frac{\xi_{ij}}{2} + \lambda(\xi_{ij})\xi_{ij}^2\right\} + \log\frac{C(e^n)}{C(e)} + \log\frac{\Gamma(a_n)}{\Gamma(a_0)} + \log\frac{\Gamma(c_n)}{\Gamma(c_0)} \\ + a_0\log b_0 + a_n(1-\frac{b_0}{b_n} - \log b_n) + c_0\log d_0 + c_n(1-\frac{d_0}{d_n} - \log d_n) \\ + \frac{1}{2}\sum_{k\leq l}^{K}\log(\sigma_\alpha^2)_{kl} + \frac{1}{2}\log|S_\beta| - \sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}\log\tau_{ik} + \frac{1}{2}\sum_{k\leq l}^{K}(\sigma_\alpha^2)_{kl}^{-1}(m_\alpha)_{kl}^2 - \frac{1}{2}m_\beta^\intercal S_\beta^{-1}m_\beta \\ + \frac{1}{2}m_\beta^\intercal\sum_{i\neq j}^{n}(Y_{ij}-\frac{1}{2})x_{ij},\end{aligned}$$

where $C(x) = \prod_{k=1}^{K}\Gamma(x_k)\Big/\Gamma\left(\sum_{k=1}^{K}x_k\right)$ *and* $\Gamma(\cdot)$ *is the gamma function.*

Residual structures While the main object of this work is to provide tools to assess the goodness of fit of a logistic regression model for networks, the considered variational algorithm also provides a natural way to estimate the residual structure ϕ. We recall that, under Model H_0, *i.e.* the network is completely explained by the covariates, the function ϕ is constant.

Still, under the alternative Model H_1, a residual structure remains, that is encoded in ϕ. As a consequence, an estimate of this function can be useful to investigate the residual structure, similarly to the residual plot classically used in a regression context. Removing the covariate effect, recall that M_K is a SBM model. Therefore, an approximate posterior mean can be derived, relying on the VBEM model averaging approach considered in [25] for SBM. Proposition 10 provides the approximate posterior mean of the function ϕ, that we propose as the network counterpart of the residual plot in regression. Note that it results from an integration over all model parameters and Models M_K.

Proposition 10 *From Proposition 1 in [25], for* $(u,v)\in[0,1]^2, u\leq v$*, the approximate posterior mean of the residual structure* ϕ *is*

$$\widehat{\mathbb{E}}\big[\phi(u,v)|Y\big] = \sum_{K\geq 1}\widehat{p}(M_K|Y)\widehat{\mathbb{E}}\big[\phi(u,v)|Y,M_K\big],$$

where

$$\widehat{\mathbb{E}}\big[\phi(u,v)|Y,M_K\big] = \sum_{k\leq l}(m_\alpha)_{kl}\left[F_{k-1,l-1}(u,v;e) - F_{k,l-1}(u,v;e) - F_{k-1,l}(u,v;e) + F_{k,l}(u,v;e)\right].$$

$F_{k,l}(u,v;e)$ *denotes the joint cdf of the Dirichlet variables* (σ_k,σ_l) *such that* $\sigma_k = \sum_{l=1}^{k}\pi_l$ *and* π *has a Dirichlet distribution Dir*(e)*.*

4 Simulation study

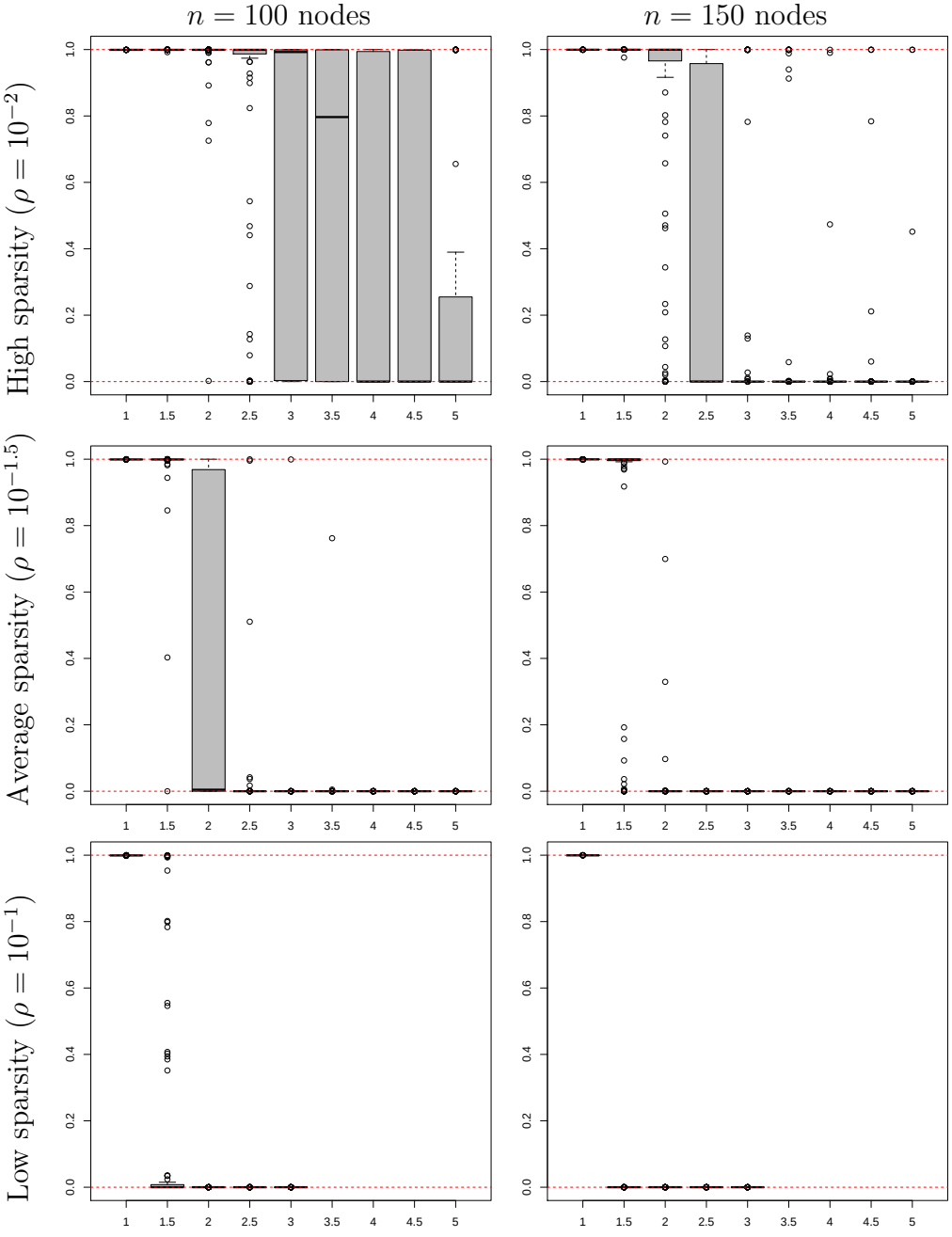
In order to assess the proposed methodology, we carried out a series of experiments on simulated data first and then on real data. In this section, we focus on the estimation of the posterior probability $\hat{p}(H_0|Y)$. We aim at evaluating the capacity of the approach to detect H_1 using toy data. Similar results were obtained for the estimated Bayes factors $\hat{B}_{01}$ and identical conclusions were drawn.

We simulated networks using Model H_1. Thus, each node is first associated to a latent position U_i sampled from a uniform distribution over the $(0, 1)$ interval. Then, a vector of covariates $x_i \in \mathbb{R}^d$ is drawn for each node, using a standardized Gaussian distribution, *i.e.* with zero mean and covariance matrix set to the identity matrix, with $d = 2$. In order to construct the covariate vector $x_{ij} \in \mathbb{R}^d$ for each edge (i, j) with $(i < j)$, we fixed $x_{ij} = x_i - x_j$. For the function $\phi(\cdot, \cdot)$, we considered a design inspired by the one proposed in [25]. In this work, the graphon function is $W(u, v) = \rho\lambda^2(uv)^{\lambda-1}$ where the parameter $\rho > 0$ controls the graph density and $\lambda > 0$ the degree concentration. For more details, we refer to [25]. Note that the maximum of the graphon function is $\rho\lambda^2$ so $\lambda < 1/\sqrt{\rho}$ must hold since $W(\cdot, \cdot)$ is a probability. In our case, the probabilities for nodes to connect are given through a logistic function $g(\cdot)$ and therefore we set $\phi(u, v) = g^{-1}\big(\rho\lambda^2(uv)^{\lambda-1}\big)$. For $\lambda = 1$, the function $\phi(\cdot, \cdot)$ is constant and so the networks are actually sampled from Model H_0. Conversely, for all $\lambda > 1$, data sets come from Model H_1. As λ increases, the residual structure, not accounted for by Model H_0, becomes sharper and thus easier to detect.

We considered networks of size $n = 100$ and $n = 150$ as well as three values for the parameter $\rho \in \{10^{-2}, 10^{-1.5}, 10^{-1}\}$ helping controlling the sparsity. Finally, we tested 9 different values of λ in $[1, 5]$. For each of the triplets (n, ρ, λ), we simulated 100 networks and we applied the methodology we propose for values of K between 1 and 10. Because the variational algorithm depends on the initialization, as any EM like procedure, for each K it was run twice and the best run was selected, such that the lower bound was maximized. Note that equal prior probabilities were given for the Models M_K $(K \geq 2)$ such that $p(H_1') = 1/2$. Moreover, we set $a_0 = b_0 = c_0 = d_0 = e_0 = 1$.

The results are presented in Figure 1. It appears that for low values of λ, the median (indicated in bold on the boxplots) of the estimated values of $\hat{p}(H_0|Y)$ is 1 and goes to 0, when λ increases, as expected. The results for the scenario with the highest sparsity ($\rho = 10^{-2}$) and $n = 100$ are unstable although the median values share this global property. Much stable results were obtained for larger networks. Interestingly, experiments can be distinguished in the way Model H_1 is detected. As soon as $\lambda > 1$, then the true model responsible for generating the data is H_1 and so the probability of Model H_0 should be lower than $1/2$. In practice, the estimated probability $\hat{p}(H_0|Y)$ is lower than $1/2$ for slightly larger values of λ. For instance, for $\rho = 10^{-1.5}$ and $n = 150$, $\hat{p}(H_0|Y) \approx 0$ for $\lambda = 2$. For $\rho = 10^{-1}$ and $n = 100$ the detection threshold appears sooner, for $\lambda = 1.5$. The experiments illustrate that H_1 is detected more easily, as the network size n increases and the sparsity parameter ρ decreases. Overall the results are encouraging with particularly low detection threshold. For $\rho = 10^{-1}$ and $n = 150$, Model H_1 is always detected when

Figure 1: Boxplots of the estimated values $\hat{p}(H_0|Y)$ of the posterior probability $p(H_0|Y)$, obtained with the variational approximations, for values of λ ranging from 1 to 5. Six scenarios considered with the number n of nodes in $\{100, 150\}$ and the sparsity parameter ρ in $\{10^{-2}, 10^{-1.5}, 10^{-1}\}$. Model H_0 is true for $\lambda = 1$ and false for $\lambda > 1$.



present ($\lambda > 1$).

5 Illustrations

We applied our approach for the analysis of two real networks from social sciences and ecology. For both studies, equal prior probabilities were given for the Models M_K ($K \geq 2$) such that $p(H_1') = 1/2$. Moreover, we set $a_0 = b_0 = c_0 = d_0 = e_0 = 1$.

5.1 Blog network

The network is made of 196 vertices and was built from a single day snapshot of political blogs extracted on 14th October 2006 [39]. Nodes correspond to blogs and an edge connect two nodes if there is an hyperlink from one blog to the other. They were annotated manually by the "Observatoire Présidentiel" project such that, for each node, labels are available. Thus, each node is associated to a political party from the left wing to the right wing and the status of the writer is also given (political analyst or not). This data set has been studied in a series of works [23, 24, 25, 39] where all the authors pointed out the crucial role of the labels in the construction of the network. The proposed methodology gives a statistical framework to decipher whether the network is fully explained by these labels built manually using expert knowledge.

We considered a set of three covariates $x_{ij} = (x_{ij}^1, x_{ij}^2, x_{ij}^3) \in \mathbb{R}^3$ artificially constructed to analyze the influence of both the political parties and the writer status. We set $x_{ij}^1 = 1$ if blogs i and j have the same labels, 0 otherwise. Moreover, $x_{ij}^2 = 1$ if one of the two blogs i and j is written by political analysts, 0 otherwise. Finally, $x_{ij}^3 = 1$ if both are written by political analysts, 0 otherwise.

We ran the variational algorithm on this data set for K between 1 and 16. For each K, the procedure was repeated 20 times and the run maximizing the lower bound was selected. We obtained a value of $\hat{p}(H_0|Y) \approx 3.6e-172$ close to zero and therefore Model H_0 was rejected. The covariates cannot explain entirely the construction of the network.

For illustration purposes, the estimation of the residual structure $g \circ \hat{\phi}$ of this data set is provided in Figure 2 ($d = 3$). In practice, we used Proposition 10 to estimate $\hat{\phi}$ and then applied $g(\cdot)$ to obtain a graphon like surface. We can observe that $g \circ \hat{\phi}$ is not constant for $d = 3$ which is coherent with H_0 being rejected. Moreover, we also give in this figure the estimated residual structure without taking the covariates into account ($d = 0$). Clearly, the shape of $g \circ \hat{\phi}$ is simpler when $d = 3$. In particular, many of the hills on the diagonal vanish when adding the covariates. Thus, the covariates help in studying and explaining parts of the network. However, they are not sufficient and some of the heterogeneity observed in the network cannot be explained by political parties and writer status.

5.2 Tree network

This data set was first introduced by [33] and further studied in [27]. We considered the tree network which describes the interactions between 51 trees where two trees interact if they share at least one common fungal parasite. Three covariates are available

Figure 2: Estimation of the blog network residual structure with and without covariates.

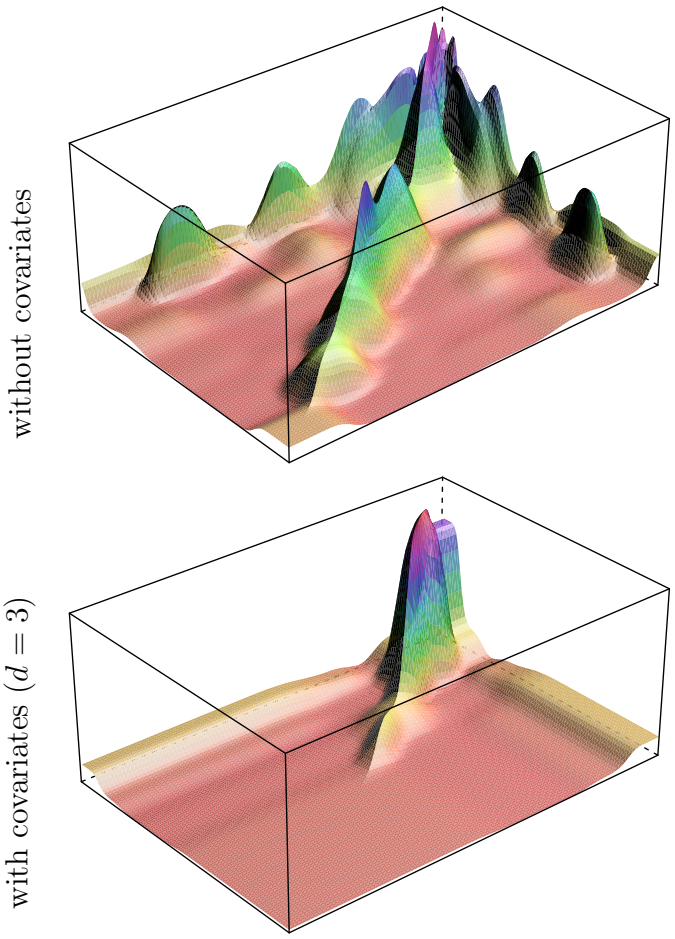
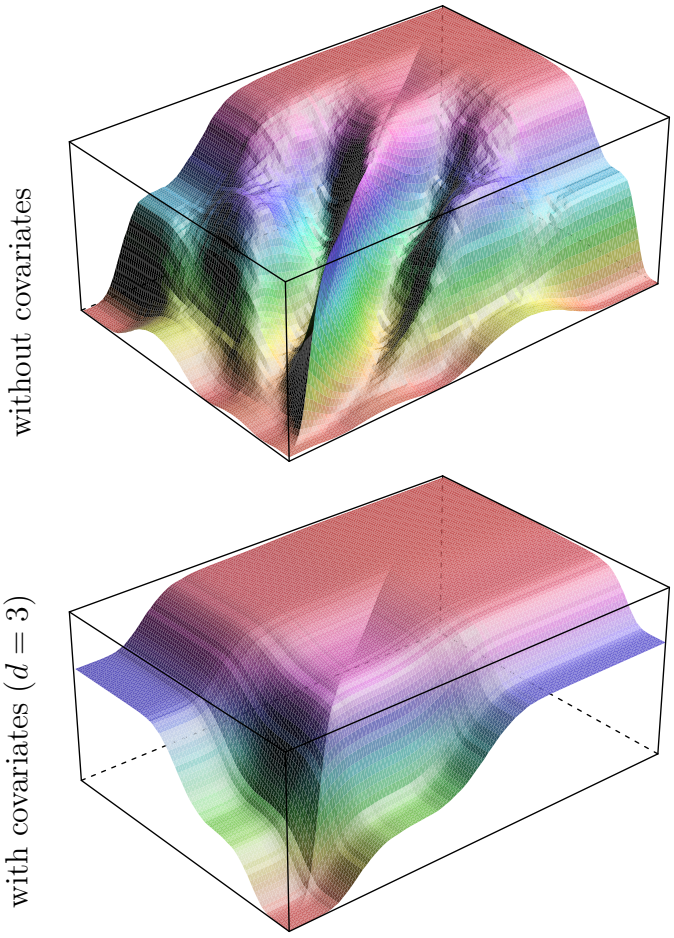


Figure 3: Estimation of the tree network residual structure with and without covariates.



characterizing the genetic, geographic, and taxonomic distances between the tree species.

As for the blog network, we applied the variational algorithm for values of K between 1 and 16. For each K, the procedure was repeated 20 times for various initializations and the best run was selected. Model H_0 was rejected with a value of $\hat{p}(H_0|Y) \approx 7.5e-153$ close to zero. Thus, the interactions between trees through common fungal parasite cannot be entirely explained by the distances available. This is consistent with a these from [27] who describe a residual heterogeneity in the valued version of this network, after taking the covariates into account.

Finally, we give the estimated residual structure $g \circ \hat{\phi}$ for this data set in Figure 3 ($d = 3$). First, we note that the structure is not constant which is coherent with H_0 being rejected. Moreover, we also provide in this figure the estimated structure without taking the covariates into account ($d = 0$). Thus, as for the blog network, we find that adding the covariates induces a simplification of $g \circ \hat{\phi}$. The extra diagonal holes vanish and the residual structure is closer to a constant function.

6 Conclusion

In this paper we proposed a framework to assess the goodness of fit of logistic models for binary networks. Thus, we added a generic term, related to the graphon function of W-graph models, to the logistic regression model. The corresponding new model was approximated with a series of models with blockwise constant residual structure. A Bayesian procedure was then considered to derive goodness-of-fit criteria. All these criteria depend on marginal likelihood terms for which we did provide estimates relying on variational approximations. The first approximation was obtained using a variational decomposition while the second involves a series of Taylor expansions. The approach was tested on toy data sets and encouraging results were obtained. Finally, it was used to analyze two real networks from social sciences and ecology. We believe the methodology has a large spectrum of applications since logistic regression models are highly used in practice to deal with covariates in binary networks.

A Appendix

A.1 Proof of Proposition 1

Let us start by showing that:

$$\log p(Y|Z,\alpha,\beta) \geq \log \sqrt{h(Z,\alpha,\beta,\xi)},$$

where ξ is an $n \times n$ positive real matrix. We use the bound on the log-logistic function introduced by [17] from Taylor expansions:

$$\log g(x) \geq \log g(\xi) + \frac{x-\xi}{2} - \lambda(\xi)(x^2-\xi^2), \forall (x,\xi) \in \mathbb{R} \times \mathbb{R}^+, \tag{7}$$

where $\lambda(\xi) = (g(\xi) - 1/2)/(2\xi)$. Note that (7) is an even function and therefore we can consider only positive values of x without loss of generality. Since

$$\log p(Y_{ij}|Z_i, Z_j, \alpha, \beta) = Y_{ij}(Z_i^\intercal \alpha Z_j + x_{ij}^\intercal \beta) + \log g(-Z_i^\intercal \alpha Z_j - x_{ij}^\intercal \beta),$$

then

$$\begin{aligned}
\log p(Y_{ij}|Z_i, Z_j, \alpha, \beta) &\geq Y_{ij}(Z_i^\intercal \alpha Z_j + x_{ij}^\intercal \beta) + \log g(\xi_{ij}) - \frac{Z_i^\intercal \alpha Z_j + x_{ij}^\intercal \beta + \xi_{ij}}{2} \\
&\quad - \lambda(\xi_{ij})((Z_i^\intercal \alpha Z_j + x_{ij}^\intercal \beta)^2 - \xi_{ij}^2) \\
&= (Y_{ij} - \frac{1}{2})(Z_i^\intercal \alpha Z_j + x_{ij}^\intercal \beta) - \frac{\xi_{ij}}{2} + \log g(\xi_{ij}) \\
&\quad - \lambda(\xi_{ij})((Z_i^\intercal \alpha Z_j + x_{ij}^\intercal \beta)^2 - \xi_{ij}^2).
\end{aligned}$$

Note that for undirected networks, the matrix ξ has to be symmetric, *i.e.* $\xi_{ij} = \xi_{ji}, \forall i \neq j$. We then have

$$\log p(Y|Z,\alpha,\beta) = \frac{1}{2}\sum_{i \neq j}^{n} \log p(Y_{ij}|Z_i, Z_j, \alpha, \beta).$$

Therefore

$$\log p(Y|Z,\alpha,\beta) \geq \log \sqrt{h(Z,\alpha,\beta,\xi)}.$$

Finally,

$$\begin{aligned}
\mathcal{L}_K(q) &= \sum_Z \int q(Z,\pi,\alpha,\beta,\gamma,\eta) \log \frac{p(Y,Z,\pi,\alpha,\beta,\gamma,\eta)}{q(Z,\pi,\alpha,\beta,\gamma,\eta)} \mathrm{d}\pi \mathrm{d}\alpha \mathrm{d}\beta \mathrm{d}\gamma \mathrm{d}\eta \\
&= \sum_Z \int q(Z,\pi,\alpha,\beta,\gamma,\eta) \log \frac{p(Y|Z,\alpha,\beta)p(Z,\pi,\alpha,\beta,\gamma,\eta)}{q(Z,\pi,\alpha,\beta,\gamma,\eta)} \mathrm{d}\pi \mathrm{d}\alpha \mathrm{d}\beta \mathrm{d}\gamma \mathrm{d}\eta \\
&\geq \sum_Z \int q(Z,\pi,\alpha,\beta,\gamma,\eta) \log \frac{\sqrt{h(Z,\alpha,\beta,\xi)}p(Z,\pi,\alpha,\beta,\gamma,\eta)}{q(Z,\pi,\alpha,\beta,\gamma,\eta)} \mathrm{d}\pi \mathrm{d}\alpha \mathrm{d}\beta \mathrm{d}\gamma \mathrm{d}\eta \\
&= \mathcal{L}_K(q;\, \xi).
\end{aligned}$$

A.2 Proof of Proposition 2

$$\begin{aligned}\log q(Z_i) &= \mathbb{E}_{Z^{\setminus i},\alpha,\beta,\pi}\left[\frac{1}{2}\log h(Z,\alpha,\beta,\xi) + \log p(Z|\pi)\right] + \text{cst} \\ &= \mathbb{E}_{Z^{\setminus i},\alpha,\beta,\pi}\left[\frac{1}{2}\sum_{i\neq j}^{n}\left\{(Y_{ij}-\frac{1}{2})Z_i^\intercal\alpha Z_j - \lambda(\xi_{ij})\left((Z_i^\intercal\alpha Z_j)^2 + 2Z_i^\intercal\alpha Z_i x_{ij}^\intercal\beta\right)\right\}\right. \\ &\qquad \left. + \sum_{i=1}^{n} Z_{ik}\log\pi_k\right] + \text{cst}.\end{aligned}$$

Note that

$$\begin{aligned}\mathbb{E}_{Z_j,\alpha}[Z_i^\intercal\alpha Z_j] &= \mathbb{E}_{Z_j,\alpha}\left[\sum_{k,l}^{K} Z_{ik}\alpha_{kl}Z_{jl}\right] \\ &= \sum_{k=1}^{K} Z_{ik}\left\{\sum_{l=1}^{K}(m_\alpha)_{kl}\tau_{jl}\right\}.\end{aligned}$$

Furthermore

$$\begin{aligned}\mathbb{E}_{Z_j,\alpha}[(Z_i^\intercal\alpha Z_j)^2] &= \mathbb{E}_{Z_j,\alpha}\left[\left(\sum_{k,l}^{K} Z_{ik}\alpha_{kl}Z_{kl}\right)^2\right] \\ &= \mathbb{E}_{Z_j,\alpha}\left[\sum_{k,k',l,l'}^{K} Z_{ik}Z_{ik'}\alpha_{kl}\alpha_{k'l'}Z_{jl}Z_{jl'}\right].\end{aligned} \tag{8}$$

Because all vectors Z_i are sampled from a multinomial distribution with parameters $(1,\pi)$, all terms $Z_{ik}Z_{ik'} = 0$ for $k \neq k^{'}$ and $Z_{ik}^2 = Z_{ik}$ in (8). Therefore

$$(Z_i^\intercal\alpha Z_j)^2 = \sum_{k,l}^{K} Z_{ik}\alpha_{kl}^2 Z_{jl}. \tag{9}$$

This leads to

$$\mathbb{E}_{Z_j,\alpha}[(Z_i^\intercal\alpha Z_j)^2] = \sum_{k=1}^{K} Z_{ik}\left\{\sum_{l=1}^{K}\mathbb{E}_{\alpha_{kl}}[\alpha_{kl}^2]\tau_{jl}\right\}.$$

Finally

$$\begin{aligned}\log q(Z_i) = \sum_{k=1}^{K} Z_{ik}\Bigg\{ &\sum_{l=1}^{K}(m_\alpha)_{kl}\frac{1}{2}\sum_{j\neq i}^{n}\Big((Y_{ij}-\frac{1}{2})-2\lambda(\xi_{ij})x_{ij}^\intercal m_\beta\Big)\tau_{jl} - \sum_{l=1}^{K}\mathbb{E}_{\alpha_{kl}}[\alpha_{kl}^2]\frac{1}{2}\sum_{j\neq i}^{n}\lambda(\xi_{ij})\tau_{jl} \\ &+ \sum_{l=1}^{K}(m_\alpha)_{lk}\frac{1}{2}\sum_{j\neq i}^{n}\Big((Y_{ji}-\frac{1}{2})-2\lambda(\xi_{ji})x_{ji}^\intercal m_\beta\Big)\tau_{jl} - \sum_{l=1}^{K}\mathbb{E}_{\alpha_{lk}}[\alpha_{lk}^2]\frac{1}{2}\sum_{j\neq i}^{n}\lambda(\xi_{ji})\tau_{jl} \\ &+ \psi(e_k^n) - \psi\Big(\sum_{l=1}^{K} e_l^n\Big)\Bigg\} + \text{cst}.\end{aligned}$$

Since $(m_\alpha)_{kl} = (m_\alpha)_{lk}$, $\mathbb{E}_{\alpha_{kl}}[\alpha_{kl}^2] = \mathbb{E}_{\alpha_{lk}}[\alpha_{lk}^2]$, $Y_{ij} = Y_{ji}$, $x_{ij} = x_{ji}$, $\xi_{ij} = \xi_{ji}$, then

$$\log q(Z_i) = \sum_{k=1}^{K} Z_{ik} \Bigg\{ \sum_{l=1}^{K} (m_\alpha)_{kl} \sum_{j \neq i}^{n} \Big((Y_{ij} - \frac{1}{2}) - 2\lambda(\xi_{ij}) x_{ij}^\intercal m_\beta \Big) \tau_{jl} - \sum_{l=1}^{K} \mathbb{E}_{\alpha_{kl}}[\alpha_{kl}^2] \sum_{j \neq i}^{n} \lambda(\xi_{ij}) \tau_{jl} + \psi(e_k^n) - \psi\Big(\sum_{l=1}^{K} e_l^n \Big) \Bigg\} + \text{cst}.$$

Therefore

$$q(Z_i) = \mathcal{M}(Z_i;\, 1, \tau_i),$$

where $\sum_{k=1}^{K} \tau_{ik} = 1$ and

$$\tau_{ik} \propto \exp \Bigg\{ \sum_{l=1}^{K} (m_\alpha)_{kl} \sum_{j \neq i}^{n} \Big((Y_{ij} - \frac{1}{2}) - 2\lambda(\xi_{ij}) x_{ij}^\intercal m_\beta \Big) \tau_{jl} - \sum_{l=1}^{K} \mathbb{E}_{\alpha_{kl}}[\alpha_{kl}^2] \sum_{j \neq i}^{n} \lambda(\xi_{ij}) \tau_{jl} + \psi(e_k^n) - \psi\Big(\sum_{l=1}^{K} e_l^n \Big) \Bigg\}.$$

A.3 Proof of Proposition 3

$$\begin{aligned} \log q(\pi) &= \mathbb{E}_Z \left[\log p(Z|\pi) + \log p(\pi) \right] + \text{cst} \\ &= \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik} \log \pi_k + \sum_{k=1}^{K} (e_0 - 1) \log \pi_k + \text{cst} \\ &= \sum_{k=1}^{K} \Big(e_0 + \sum_{i=1}^{n} \tau_{ik} - 1 \Big) \log \pi_k + \text{cst}. \end{aligned}$$

Therefore

$$q(\pi) = \text{Dir}(\pi;\, e^n),$$

where $e_k^n = e_0 + \sum_{i=1}^{n} \tau_{ik}, \forall k \in \{1, \ldots, K\}$.

A.4 Proof of Proposition 4

$$\begin{aligned} \log q(\beta) &= \mathbb{E}_{Z,\alpha,\eta} \left[\frac{1}{2} \log h(Z, \alpha, \beta, \xi) + \log p(\beta|\eta) \right] + \text{cst} \\ &= \mathbb{E}_{Z,\alpha,\eta} \left[\frac{1}{2} \sum_{i \neq j}^{n} \Big\{ (Y_{ij} - 1/2) x_{ij}^\intercal \beta - \lambda(\xi_{ij}) \Big((x_{ij}^\intercal \beta)^2 + 2 Z_i^\intercal \alpha Z_j x_{ij}^\intercal \beta \Big) \Big\} - \frac{\eta}{2} \beta^\intercal \beta \right] + \text{cst} \\ &= \beta^\intercal \left\{ \frac{1}{2} \sum_{i \neq j}^{n} \Big(Y_{ij} - \frac{1}{2} - 2\lambda(\xi_{ij}) \tau_i^\intercal m_\alpha \tau_j \Big) x_{ij} \right\} - \frac{1}{2} \beta^\intercal \left\{ \frac{c_n}{d_n} I_d + \sum_{i \neq j}^{n} \lambda(\xi_{ij}) x_{ij} x_{ij}^\intercal \right\} \beta + \text{cst}. \end{aligned}$$

Therefore

$$q(\beta) = \mathcal{N}(\beta;\, m_\beta, S_\beta),$$

where

$$S_\beta^{-1} = \frac{c_n}{d_n} I_d + \sum_{i \neq j}^{n} \lambda(\xi_{ij}) x_{ij} x_{ij}^\intercal,$$

and

$$m_\beta = S_\beta \frac{1}{2} \sum_{i \neq j}^{n} \left(Y_{ij} - \frac{1}{2} - 2\lambda(\xi_{ij}) \tau_i^\intercal m_\alpha \tau_j \right) x_{ij}.$$

A.5 Proof of Proposition 5

$$\begin{aligned}
\log q(\gamma) &= \mathbb{E}_\alpha \left[\log p(\alpha|\gamma) + \log p(\gamma) \right] + \text{cst} \\
&= \mathbb{E}_\alpha \left[\sum_{k \leq l}^{K} \frac{1}{2} \log(\gamma) - \sum_{k \leq l}^{K} \frac{\gamma}{2} \alpha_{kl}^2 \right] + (a_0 - 1) \log \gamma - b_0 \gamma + \text{cst} \\
&= (a_0 + \frac{K(K+1)}{4} - 1) \log \gamma - \left(b_0 + \frac{1}{2} \sum_{k \leq l}^{K} \mathbb{E}_{\alpha_{kl}}[\alpha_{kl}^2] \right) \gamma + \text{cst}.
\end{aligned}$$

Therefore

$$q(\gamma) = \text{Gam}(\gamma;\, a_n, b_n),$$

where $a_n = a_0 + \frac{K(K+1)}{4}$ and $b_n = b_0 + \frac{1}{2} \sum_{k \leq l}^{K} \mathbb{E}_{\alpha_{kl}}[\alpha_{kl}^2]$.

A.6 Proof of Proposition 6

$$\begin{aligned}
\log q(\eta) &= \mathbb{E}_\beta \left[\log p(\beta|\eta) + \log p(\eta) \right] + \text{cst} \\
&= \mathbb{E}_\beta \left[\frac{d}{2} \log \eta - \frac{\eta}{2} \beta^\intercal \beta \right] + (c_0 - 1) \log \eta - d_0 \eta + \text{cst} \\
&= (c_0 + \frac{d}{2} - 1) \log \eta - (d_0 + \frac{1}{2} \text{Tr}(S_\beta) + \frac{1}{2} m_\beta^\intercal m_\beta) \eta + \text{cst}.
\end{aligned}$$

Therefore

$$q(\eta) = \text{Gam}(\eta;\, c_n, d_n),$$

where $c_n = c_0 + \frac{d}{2}$ and $d_n = d_0 + \frac{1}{2} \text{Tr}(S_\beta) + \frac{1}{2} m_\beta^\intercal m_\beta$.

A.7 Proof of Proposition 7

$$\begin{aligned}\log q(\alpha) &= \mathbb{E}_{Z,\beta,\gamma}\left[\frac{1}{2}\log h(Z,\alpha,\beta,\xi) + \log p(\alpha|\gamma)\right] + \text{cst} \\ &= \mathbb{E}_{Z,\beta,\gamma}\left[\frac{1}{2}\sum_{i\neq j}^{n}\left\{(Y_{ij}-\frac{1}{2})Z_i^\intercal \alpha Z_j - \lambda(\xi_{ij})\Big((Z_i^\intercal\alpha Z_j)^2 + 2Z_i^\intercal\alpha Z_j x_{ij}^\intercal\beta\Big)\right\}\right. \qquad (10)\\ &\qquad \left. - \sum_{k\leq l}^{K}\frac{\gamma}{2}\alpha_{kl}^2\right] + \text{cst}.\end{aligned}$$

We have $Z_i^\intercal\alpha Z_j = \sum_{k,l}^K Z_{ik}\alpha_{kl}Z_{kl}$ and $(Z_i^\intercal\alpha Z_j)^2 = \sum_{k,l}^K Z_{ik}\alpha_{kl}^2 Z_{kl} = Z_i^\intercal A Z_j$ (see Eq. 9) with A the $K\times K$ matrix such that $A_{kl}=\alpha_{kl}^2$. Moreover, any expression of the form $(1/2)\sum_{i\neq j}^n c_{ij}Z_i^\intercal B Z_j$ where B is a symmetric $K\times K$ matrix and $c_{ij}=c_{ji}$ can be written

$$\begin{aligned}\frac{1}{2}\sum_{i\neq j}^{n} c_{ij}Z_i^\intercal B Z_j &= \frac{1}{2}\sum_{i\neq j}^{n} c_{ij}\sum_{k,l}^{K} Z_{ik}B_{kl}Z_{jl} \\ &= \frac{1}{2}\sum_{i\neq j}^{n} c_{ij}\left(\sum_{k=1}^{K} Z_{ik}B_{kk}Z_{jk} + \sum_{k<l}^{K} Z_{ik}B_{kl}Z_{jl} + \sum_{k<l}^{K} Z_{jk}B_{lk}Z_{il}\right) \\ &= \sum_{k=1}^{K} B_{kk}\frac{1}{2}\sum_{i\neq j}^{n} c_{ij}Z_{ik}Z_{jk} + \sum_{k<l}^{K} B_{kl}\left(\frac{1}{2}\sum_{i\neq j}^{n} c_{ij}Z_{ik}Z_{jl} + \frac{1}{2}\sum_{i\neq j}^{n} c_{ij}Z_{jk}Z_{il}\right).\end{aligned}$$

By exchanging the role of i and j in the sum of the last term and since $c_{ij}=c_{ji}$, we obtain

$$\frac{1}{2}\sum_{i\neq j}^{n} c_{ij}Z_i^\intercal B Z_j = \sum_{k=1}^{K} B_{kk}\frac{1}{2}\sum_{i\neq j}^{n} c_{ij}Z_{ik}Z_{jk} + \sum_{k<l}^{K} B_{kl}\sum_{i\neq j}^{n} c_{ij}Z_{ik}Z_{jl}. \qquad (11)$$

Using (11) in (10) leads to

$$\begin{aligned}\log q(\alpha) = &\sum_{k=1}^{K}\alpha_{kk}\sum_{i\neq j}^{n}\left(\frac{1}{2}(Y_{ij}-\frac{1}{2}) - \lambda(\xi_{ij})x_{ij}^\intercal m_\beta\right)\tau_{ik}\tau_{jk} \\ &+ \sum_{k<l}^{K}\alpha_{kl}\sum_{i\neq j}^{n}\left((Y_{ij}-\frac{1}{2}) - 2\lambda(\xi_{ij})x_{ij}^\intercal m_\beta\right)\tau_{ik}\tau_{jl} \\ &- \sum_{k=1}^{K}\alpha_{kk}^2\sum_{i\neq j}^{n}\frac{1}{2}\lambda(\xi_{ij})\tau_{ik}\tau_{jk} - \sum_{k<l}^{K}\alpha_{kl}^2\sum_{i\neq j}^{n}\lambda(\xi_{ij})\tau_{ik}\tau_{jl} \\ &- \sum_{k\leq l}^{K}\frac{a_n}{2b_n}\alpha_{kl}^2.\end{aligned}$$

Therefore

$$q(\alpha) = \prod_{k\neq l}^{K}\mathcal{N}\left(\alpha_{kl}; (m_\alpha)_{kl}, (\sigma_\alpha^2)_{kl}\right),$$

where

$$(\sigma_\alpha^2)_{kk}^{-1} = \frac{a_n}{b_n} + \sum_{i\neq j}^{n} \lambda(\xi_{ij})\tau_{ik}\tau_{jk}, \forall k,$$

$$(\sigma_\alpha^2)_{kl}^{-1} = \frac{a_n}{b_n} + 2\sum_{i\neq j}^{n} \lambda(\xi_{ij})\tau_{ik}\tau_{jl}, \forall k \neq l,$$

$$(m_\alpha)_{kk} = (\sigma_\alpha^2)_{kk} \sum_{i\neq j}^{n} \left(\frac{1}{2}(Y_{ij} - \frac{1}{2}) - \lambda(\xi_{ij}) x_{ij}^\intercal m_\beta\right) \tau_{ik}\tau_{jk}$$

$$(m_\alpha)_{kl} = (\sigma_\alpha^2)_{kl} \sum_{i\neq j}^{n} \left((Y_{ij} - \frac{1}{2}) - 2\lambda(\xi_{ij}) x_{ij}^\intercal m_\beta\right) \tau_{ik}\tau_{jl}.$$

A.8 Proof of Proposition 9

$\mathcal{L}_K(q;\xi) = \mathbb{E}_{Z,\pi,\alpha,\beta,\gamma,\eta}\left[\log\sqrt{h(Z,\alpha,\beta,\xi)} + \log p(Z,\pi,\alpha,\beta,\gamma,\eta)\right] - \mathbb{E}_{Z,\pi,\alpha,\beta,\gamma,\eta}\left[\log q(Z,\pi,\alpha,\beta,\gamma,\eta)\right]$

$$\begin{aligned}
\mathcal{L}_K(q;\,\xi) = \frac{1}{2}\sum_{i\neq j}^{n}\Bigg\{ & (Y_{ij} - \frac{1}{2})(\mathbb{E}_{Z_i,Z_j,\alpha}[Z_i^\intercal \alpha Z_j] + x_{ij}^\intercal \mathbb{E}_\beta[\beta]) + \log g(\xi_{ij}) - \frac{\xi_{ij}}{2} \\
& - \lambda(\xi_{ij})\left(\mathbb{E}_{Z_i,Z_j,\alpha}[(Z_i^\intercal \alpha Z_j)^2] + 2\mathbb{E}_{Z_i,Z_j,\alpha}[Z_i^\intercal \alpha Z_j] x_{ij}^\intercal \mathbb{E}_\beta[\beta] + \mathbb{E}_\beta[(x_{ij}^\intercal \beta)^2] - \xi_{ij}^2\right)\Bigg\} \\
& + \sum_{i=1}^{n}\sum_{k=1}^{K} \mathbb{E}_{Z_i}[Z_{ik}]\mathbb{E}_\pi[\log\pi_k] - \log C(e) + \sum_{k=1}^{K}(e_0 - 1)\mathbb{E}_\pi[\log\pi_k] - \frac{K(K+1)}{4}\log(2\pi) \\
& + \frac{K(K+1)}{4}\mathbb{E}_\gamma[\log\gamma] - \frac{\mathbb{E}_\gamma[\gamma]}{2}\sum_{k\leq l}^{K}\mathbb{E}_{\alpha_{kl}}[\alpha_{kl}^2] - \frac{d}{2}\log(2\pi) + \frac{d}{2}\mathbb{E}_\eta[\log\eta] - \frac{\mathbb{E}_\eta[\eta]}{2}\mathbb{E}_\beta[\beta^\intercal\beta] \\
& - \log\Gamma(a_0) + a_0\log b_0 + (a_0 - 1)\mathbb{E}_\gamma[\log\gamma] - b_0\mathbb{E}_\gamma[\gamma] - \log\Gamma(c_0) \\
& + c_0\log d_0 + (c_0 - 1)\mathbb{E}_\eta[\log\eta] - d_0\mathbb{E}_\eta[\eta] - \sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{E}_{Z_i}[Z_{ik}]\log\tau_{ik} + \log C(e^n) \\
& - \sum_{k=1}^{K}(e_k^n - 1)\mathbb{E}_\pi[\log\pi_k] + \frac{K(K+1)}{4}\log(2\pi) + \frac{1}{2}\sum_{k\leq l}^{K}\log(\sigma_\alpha^2)_{kl} + \frac{1}{2}\sum_{k\leq l}^{K}(\sigma_\alpha^2)_{kl}^{-1}\mathbb{E}_{\alpha_{kl}}[\alpha_{kl}^2] \\
& - \sum_{k\leq l}^{K}(\sigma_\alpha^2)_{kl}^{-1}\mathbb{E}_{\alpha_{kl}}[\alpha_{kl}](m_\alpha)_{kl} + \frac{1}{2}\sum_{k\leq l}^{K}(\sigma_\alpha^2)_{kl}^{-1}(m_\alpha)_{kl}^2 + \frac{d}{2}\log(2\pi) + \frac{1}{2}\log|S_\beta| \\
& + \frac{1}{2}\mathbb{E}_\beta[\beta^\intercal S_\beta^{-1}\beta] - \mathbb{E}_\beta[\beta]^\intercal S_\beta^{-1} m_\beta + \frac{1}{2}m_\beta S_\beta^{-1} m_\beta + \log\Gamma(a_n) - a_n\log b_n \\
& - (a_n - 1)\mathbb{E}_\gamma[\log\gamma] + b_n\mathbb{E}_\gamma[\gamma] + \log\Gamma(c_n) - c_n\log d_n - (c_n - 1)\mathbb{E}_\eta[\log\eta] \\
& + d_n\mathbb{E}_\eta[\eta], \quad (12)
\end{aligned}$$

where $C(x) = \frac{\prod_{k=1}^{K}\Gamma(x_k)}{\Gamma(\sum_{k=1}^{K}x_k)}$ and $\Gamma(\cdot)$ is the gamma function. The terms in $\mathbb{E}_\gamma[\log\gamma]$, $\mathbb{E}_\eta[\log\eta]$, $\mathbb{E}_\pi[\log\pi]$ and $\log(2\pi)$ do simplify in (12) after the VBEM update step. This leads to

$$\begin{aligned}
\mathcal{L}_K(q;\,\xi) = \frac{1}{2}\sum_{i\neq j}^{n}\Bigg\{ & (Y_{ij}-\frac{1}{2})(\mathbb{E}_{Z_i,Z_j,\alpha}[Z_i^\intercal\alpha Z_j] + x_{ij}^\intercal\mathbb{E}_\beta[\beta]) + \log g(\xi_{ij}) - \frac{\xi_{ij}}{2} \\
& - \lambda(\xi_{ij})\left(\mathbb{E}_{Z_i,Z_j,\alpha}[(Z_i^\intercal\alpha Z_j)^2] + 2\mathbb{E}_{Z_i,Z_j,\alpha}[Z_i^\intercal\alpha Z_j]x_{ij}^\intercal\mathbb{E}_\beta[\beta] + \mathbb{E}_\beta[(x_{ij}^\intercal\beta)^2] - \xi_{ij}^2\right)\Bigg\} \\
& - \log C(e) - \frac{a_n}{2b_n}\sum_{k\leq l}^{K}\mathbb{E}_{\alpha_{kl}}[\alpha_{kl}^2] - \frac{c_n}{2d_n}\mathrm{Tr}(S_\beta + m_\beta m_\beta^\intercal) \\
& - \log\Gamma(a_0) + a_0\log b_0 - b_0\frac{a_n}{b_n} - \log\Gamma(c_0) + c_0\log d_0 - d_0\frac{c_n}{d_n} \\
& - \sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}\log\tau_{ik} + \log C(e^n) + \frac{1}{2}\sum_{k\leq l}^{K}\log(\sigma_\alpha^2)_{kl} + \frac{1}{2}\sum_{k\leq l}^{K}(\sigma_\alpha^2)_{kl}^{-1}\mathbb{E}_{\alpha_{kl}}[\alpha_{kl}^2] \\
& - \frac{1}{2}\sum_{k\leq l}^{K}(\sigma_\alpha^2)_{kl}^{-1}(m_\alpha)_{kl}^2 + \frac{1}{2}\log|S_\beta| + \frac{1}{2}\mathrm{Tr}\big(S_\beta^{-1}(S_\beta + m_\beta(m_\beta)^\intercal)\big) \\
& - \frac{1}{2}m_\beta^\intercal S_\beta^{-1}m_\beta + \log\Gamma(a_n) - a_n\log b_n + b_n\frac{a_n}{b_n} + \log\Gamma(c_n) \\
& - c_n\log d_n + d_n\frac{c_n}{d_n}.
\end{aligned}$$

Moreover, using (11), note that

$$\begin{aligned}
\frac{1}{2}\sum_{i\neq j}^{n}(Y_{ij}-\frac{1}{2})\mathbb{E}_{Z_i,Z_j,\alpha}[Z_i^\intercal\alpha Z_j] = & \sum_{k=1}^{K}\mathbb{E}_{\alpha_{kk}}[\alpha_{kk}]\frac{1}{2}\sum_{i\neq j}^{n}(Y_{ij}-\frac{1}{2})\tau_{ik}\tau_{jk} \\
& + \sum_{k<l}^{K}\mathbb{E}_{\alpha_{kl}}[\alpha_{kl}]\sum_{i\neq j}^{n}(Y_{ij}-\frac{1}{2})\tau_{ik}\tau_{jl},
\end{aligned}$$

$$\begin{aligned}
\frac{1}{2}\sum_{i\neq j}^{n}2\lambda(\xi_{ij})\mathbb{E}_{Z_i,Z_j,\alpha}[Z_i^\intercal\alpha Z_j]x_{ij}^\intercal\mathbb{E}_\beta[\beta] = & \sum_{k=1}^{K}\mathbb{E}_{\alpha_{kk}}[\alpha_{kk}]\sum_{i\neq j}^{n}\lambda(\xi_{ij})x_{ij}^\intercal\mathbb{E}_\beta[\beta]\tau_{ik}\tau_{jk} \\
& + \sum_{k<l}^{K}\mathbb{E}_{\alpha_{kl}}[\alpha_{kl}]2\sum_{i\neq j}^{n}\lambda(\xi_{ij})x_{ij}^\intercal\mathbb{E}_\beta[\beta]\tau_{ik}\tau_{jl}.
\end{aligned}$$

Using (9) and (11),

$$\frac{1}{2}\sum_{i\neq j}^{n}\lambda(\xi_{ij})\mathbb{E}_{Z_i,Z_j,\alpha}[(Z_i^\intercal\alpha Z_j)^2] = \sum_{k=1}^{K}\mathbb{E}_{\alpha_{kk}}[\alpha_{kk}^2]\frac{1}{2}\sum_{i\neq j}^{n}\lambda(\xi_{ij})\tau_{ik}\tau_{jk} + \sum_{k<l}^{K}\mathbb{E}_{\alpha_{kl}}[\alpha_{kl}^2]\sum_{i\neq j}^{n}\lambda(\xi_{ij})\tau_{ik}\tau_{jl}$$

Finally

$$\begin{aligned}\mathbb{E}_\beta[(x_{ij}^\intercal\beta)^2] &= \mathbb{E}_\beta\left[x_{ij}^\intercal\beta x_{ij}^\intercal\beta\right]\\ &= \mathbb{E}_\beta\left[x_{ij}^\intercal\beta\beta^\intercal x_{ij}\right]\\ &= \mathrm{Tr}\left(x_{ij}x_{ij}^\intercal\mathbb{E}_\beta[\beta\beta^\intercal]\right)\\ &= \mathrm{Tr}\left(x_{ij}x_{ij}^\intercal(S_\beta + m_\beta m_\beta^\intercal)\right).\end{aligned}$$

Therefore

$$\begin{aligned}\mathcal{L}_K(q;\,\xi) = {}& \frac{1}{2}\sum_{i\neq j}^{n}\left\{\log g(\xi_{ij}) - \frac{\xi_{ij}}{2} + \lambda(\xi_{ij})\xi_{ij}^2\right\} + \log\frac{C(e^n)}{C(e)} + \log\frac{\Gamma(a_n)}{\Gamma(a_0)} + \log\frac{\Gamma(c_n)}{\Gamma(c_0)}\\ &+ a_0\log b_0 + a_n(1-\frac{b_0}{b_n} - \log b_n) + c_0\log d_0 + c_n(1-\frac{d_0}{d_n} - \log d_n)\\ &+ \frac{1}{2}\sum_{k\leq l}^{K}\log(\sigma_\alpha^2)_{kl} + \frac{1}{2}\log|S_\beta| - \sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}\log\tau_{ik} - \frac{1}{2}\sum_{k\leq l}^{K}(\sigma_\alpha^2)_{kl}^{-1}(m_\alpha)_{kl}^2 - \frac{1}{2}m_\beta^\intercal S_\beta^{-1}m_\beta\\ &+ \sum_{k=1}^{K}(m_\alpha)_{kk}\sum_{i\neq j}^{n}\left(\frac{1}{2}(Y_{ij}-\frac{1}{2}) - \lambda(\xi_{ij})x_{ij}^\intercal m_\beta\right)\tau_{ik}\tau_{jk}\\ &+ \sum_{k<l}^{K}(m_\alpha)_{kl}\sum_{i\neq j}^{n}\left((Y_{ij}-\frac{1}{2}) - 2\lambda(\xi_{ij})x_{ij}^\intercal m_\beta\right)\tau_{ik}\tau_{jl}\\ &- \sum_{k=1}^{K}\mathbb{E}_{\alpha_{kk}}[\alpha_{kk}^2]\frac{1}{2}\left(\frac{a_n}{b_n} - (\sigma_\alpha^2)_{kk}^{-1} + \sum_{i\neq j}^{n}\lambda(\xi_{ij})\tau_{ik}\tau_{jk}\right)\\ &- \sum_{k<l}^{K}\mathbb{E}_{\alpha_{kl}}[\alpha_{kl}^2]\frac{1}{2}\left(\frac{a_n}{b_n} - (\sigma_\alpha^2)_{kl}^{-1} + 2\sum_{i\neq j}^{n}\lambda(\xi_{ij})\tau_{ik}\tau_{jl}\right)\\ &+ \frac{1}{2}m_\beta^\intercal\sum_{i\neq j}^{n}(Y_{ij}-\frac{1}{2})x_{ij}\\ &- \frac{1}{2}\mathrm{Tr}\left(\left(2\sum_{i\neq j}^{n}\lambda(\xi_{ij})x_{ij}x_{ij}^\intercal + \frac{c_n}{d_n}I_d - S_\beta^{-1}\right)(S_\beta + m_\beta m_\beta^\intercal)\right).\end{aligned}$$

Finally, since the terms at the fourth and fifth line correspond exactly to $\sum_{k\leq l}^{K}(m_\alpha)_{kl}(\sigma_\alpha^2)_{kl}^{-1}(m_\alpha)_{kl}$, and after the VBEM update step

$$\begin{aligned}\mathcal{L}_K(q;\,\xi) = \frac{1}{2}\sum_{i\neq j}^{n}\left\{\log g(\xi_{ij}) - \frac{\xi_{ij}}{2} + \lambda(\xi_{ij})\xi_{ij}^2\right\} + \log\frac{C(e^n)}{C(e)} + \log\frac{\Gamma(a_n)}{\Gamma(a_0)} + \log\frac{\Gamma(c_n)}{\Gamma(c_0)} \\ + a_0\log b_0 + a_n(1-\frac{b_0}{b_n} - \log b_n) + c_0\log d_0 + c_n(1-\frac{d_0}{d_n} - \log d_n) \\ + \frac{1}{2}\sum_{k\leq l}^{K}\log(\sigma_\alpha^2)_{kl} + \frac{1}{2}\log|S_\beta| - \sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}\log\tau_{ik} + \frac{1}{2}\sum_{k\leq l}^{K}(\sigma_\alpha^2)_{kl}^{-1}(m_\alpha)_{kl}^2 - \frac{1}{2}m_\beta^\intercal S_\beta^{-1}m_\beta \\ + \frac{1}{2}m_\beta^\intercal\sum_{i\neq j}^{n}(Y_{ij}-\frac{1}{2})x_{ij}.\end{aligned}$$

A.9 Proof of Proposition 8

Keeping only the terms that do depend on ξ_{ij} in (12), the lower bound is given by

$$\begin{aligned}\mathcal{L}_K(q;\,\xi) &= \frac{1}{2}\sum_{i\neq j}^{n}\Bigg\{\log g(\xi_{ij}) - \frac{\xi_{ij}}{2} - \lambda(\xi_{ij})\Big(\mathbb{E}_{Z_i,Z_j,\alpha}[(Z_i^\intercal\alpha Z_j)^2] + 2\mathbb{E}_{Z_i,Z_j,\alpha}[Z_i^\intercal\alpha Z_j]x_{ij}^\intercal\mathbb{E}_\beta[\beta] \\ &\quad + \mathbb{E}_\beta[(x_{ij}^\intercal\beta)^2] - \xi_{ij}^2\Big)\Bigg\} + \text{cst} \\ &= \frac{1}{2}\sum_{i\neq j}^{n}\Bigg\{\log g(\xi_{ij}) - \frac{\xi_{ij}}{2} - \lambda(\xi_{ij})\Big(\sum_{k,l}^{K}\tau_{ik}\tau_{jl}\mathbb{E}_{\alpha_{kl}}[\alpha_{kl}^2] + 2\sum_{k,l}^{K}\tau_{ik}\tau_{jl}(m_\alpha)_{kl}x_{ij}^\intercal m_\beta \\ &\quad + \text{Tr}(x_{ij}x_{ij}^\intercal(S_\beta + m_\beta m_\beta^\intercal)) - \xi_{ij}^2\Big)\Bigg\} + \text{cst}.\end{aligned}$$

The partial derivative of the lower bound with respect to ξ_{ij} is

$$\begin{aligned}\frac{\partial\mathcal{L}_K}{\partial\xi_{ij}}(q;\,\xi) = g(-\xi_{ij}) - \frac{1}{2} - \lambda'(\xi_{ij})\Big(\sum_{k,l}^{K}\tau_{ik}\tau_{jl}\mathbb{E}_{\alpha_{kl}}[\alpha_{kl}^2] + 2\sum_{k,l}^{K}\tau_{ik}\tau_{jl}(m_\alpha)_{kl}x_{ij}^\intercal m_\beta \\ + \text{Tr}(x_{ij}x_{ij}^\intercal(S_\beta + m_\beta m_\beta^\intercal)) - \xi_{ij}^2\Big) + 2\lambda(\xi_{ij})\xi_{ij},\end{aligned}$$

and

$$\begin{aligned}\frac{\partial\mathcal{L}_K}{\partial\xi_{ij}}(q;\,\xi) = -\lambda'(\xi_{ij})\Big(\sum_{k,l}^{K}\tau_{ik}\tau_{jl}\mathbb{E}_{\alpha_{kl}}[\alpha_{kl}^2] + 2\sum_{k,l}^{K}\tau_{ik}\tau_{jl}(m_\alpha)_{kl}x_{ij}^\intercal m_\beta \\ + \text{Tr}(x_{ij}x_{ij}^\intercal(S_\beta + m_\beta m_\beta^\intercal)) - \xi_{ij}^2\Big). \quad (13)\end{aligned}$$

Finally, $\lambda(\xi_{ij})$ is a strictly decreasing function for positive values of ξ_{ij} . Thus, $\lambda^{'}(\xi_{ij}) \neq 0$ and therefore if we set (13) to zero, we obtain

$$\xi_{ij}^2 = \sum_{k,l}^{K} \tau_{ik}\tau_{jl}\mathbb{E}_{\alpha_{kl}}[\alpha_{kl}^2] + 2\sum_{k,l}^{K} \tau_{ik}\tau_{jl}(m_\alpha)_{kl}x_{ij}^\intercal m_\beta + \mathrm{Tr}(x_{ij}x_{ij}^\intercal(S_\beta + m_\beta m_\beta^\intercal)).$$

Note that $\xi_{ij} = \xi_{ji}$.

References

[1] Edoardo M Airoldi, Thiago B Costa, and Stanley H Chan. Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Advances in Neural Information Processing Systems*, pages 692–700, 2013.

[2] R. Albert and A.L. Barabási. Statistical mechanics of complex networks. *Modern Physics*, 74:47–97, 2002.

[3] D. Asta and C. R. Shalizi. Geometric network comparison. Technical report, arXiv:1411.1350v1, 2014.

[4] A.L. Barabási and Z.N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Rev. Genet*, 5:101–113, 2004.

[5] M.J. Beal and Z. Ghahramani. The variational bayesian em algorithm for incomplete data: with application to scoring graphical model structures. In JM Bernardo, MJ Bayarri, JO Berger, AP Dawid, D Heckerman, AFM Smith, and M (eds) West, editors, *Bayesian Statistics 7: Proceedings of the 7th Valencia International Meeting*, page 453, 2002.

[6] C.M. Bishop. *Pattern recognition and machine learning*. Springer-Verlag, 2006.

[7] C.M. Bishop and M. Svensén. Bayesian hierarchical mixtures of experts. In *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence*, pages 57–64. U. Kjaerulff and C. Meek, 2003.

[8] F. Caron and A. Doucet. Sparse bayesian nonparametric regression. In *Proceedings of the 25th International Conference on Machine Learning*, 2008.

[9] S. Chatterjee. Matrix estimation by Universal Singular Value Thresholding. *The Annals of Statistics*, 43(1):177–214, 2015.

[10] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood for incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, B39:1–38, 1977.

[11] P. Diaconis and S. Janson. Graph limits and exchangeable random graphs. *Rend. Mat. Appl.*, 7(28):33–61, 2008.

[12] C. Ducruet. Network diversity and maritime flows. *Journal of Transport Geography*, 30:77–88, 2013.

[13] A. Goldenberg, A.X. Zheng, S.E. Fienberg, and E.M. Airoldi. A survey of statistical network models. *Foundations and Trends in Machine Learning*, 2(2):129–233, 2010.

[14] Mark S Handcock, Adrian E Raftery, and Jeremy M Tantrum. Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–354, 2007.

[15] P.D. Hoff. Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in Neural Information Processing Systems*, pages 657–664, 2008.

[16] Peter D Hoff, Adrian E Raftery, and Mark S Handcock. Latent space approaches to social network analysis. *Journal of the american Statistical association*, 97(460):1090–1098, 2002.

[17] T.S. Jaakkola and M.I. Jordan. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10:25–37, 2000.

[18] H. Jeffreys. An invariant form for the prior probability in estimations problems. In *Proceedings of the Royal Society of London. Series A*, volume 186, pages 453–461, 1946.

[19] Y. Jernite, P. Latouche, C. Bouveyron, P. Rivera, L. Jegou, and S. Lamassé. The random subgraph model for the analysis of an acclesiastical network in merovingian gaul. *Annals of Applied Statistics*, 8(1):377–405, 2014.

[20] O. Kallenberg. Multivariate sampling and the estimation problem for exchangeable arrays. *Journal of Theoretical Probability*, 12(3):859–883, 1999.

[21] R. E Kass and A. E Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

[22] V. Lacroix, C.G. Fernandes, and M.-F. Sagot. Motif search in graphs:application to metabolic networks. *Transactions in Computational Biology and Bioinformatics*, 3:360–368, 2006.

[23] P. Latouche, E Birmelé, and C. Ambroise. Overlapping stochastic block models with application to the french political blogosphere. *Annals of Applied Statistics*, 5(1):309–336, 2011.

[24] P. Latouche, E Birmelé, and C. Ambroise. Model selection in overlapping stochastic block models. *Electronic Journal of Statistics*, 8(1):762–794, 2014.

[25] P. Latouche and S Robin. Bayesian model averaging of stochastic block models to estimate the graphon function and motif frequencies in a W-graph model. Technical report, arXiv:1310.6150, 2013.

[26] L. Lovász and B. Szegedy. Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B*, 96(6):933 – 957, 2006.

[27] M. Mariadassou, S. Robin, and C. Vacher. Uncovering latent structure in valued graphs: a variational approach. *Annals of Applied Statistics*, 4(2), 2010.

[28] C. Matias and S. Robin. Modeling heterogenity in random graphs through latent space models: a selective review. *Esaim Prooceedings and Surveys*, 47:55–74, 2014.

[29] Mark E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.

[30] K. Nowicki and T.A.B. Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96:1077–1087, 2001.

[31] G. Palla, L. Lovasz, and T. Vicsek. Multifractal network generator. *Proc. Natl. Acad. Sci. U.S.A.*, 107(17):7640–7645, Apr 2010.

[32] T.A.B. Snijders and K. Nowicki. Estimation and prediction for stochastic blockstructures for graphs with latent block structure. *Journal of Classification*, 14:75–100, 1997.

[33] C. Vacher, D. Piou, and M.-L. Desprez-Loustau. Architecture of an antagonistic tree/fungus network: The asymmetric influence of past evolutionary history. *PLoS ONE*, 3(3):1740, 2008. `e1740. doi:10.1371/journal.pone.0001740`.

[34] N. Villa, F. Rossi, and Q.D. Truong. Mining a medieval social network by kernel som and related methods. *Arxiv preprint arXiv:0805.1374*, 2008.

[35] Stevenn Volant, Marie-Laure Martin Magniette, and Stéphane Robin. Variational bayes approach for model aggregation in unsupervised classification with markovian dependency. *Comput. Statis. & Data Analysis*, 56(8):2375 – 2387, 2012.

[36] Y.J. Wang and G.Y. Wong. Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, 82:8–19, 1987.

[37] D.J. Watts and S.H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393:440–442, 1998.

[38] P. J. Wolfe and S. C. Olhede. Nonparametric graphon estimation. Technical report, arXiv:1309.5936, 2013.

[39] H. Zanghi, C. Ambroise, and V. Miele. Fast online graph clustering via erdös-rényi mixture. *Pattern Recognition*, 41(12):3592–3599, 2008.

[40] H. Zanghi, S. Volant, and C. Ambroise. Clustering based on random graph model embedding vertex features. *Pattern Recognition Letters*, 31(9):830–836, 2010.