Strong consistency of Krichevsky-Trofimov estimator for the number of communities in the Stochastic Block Model*

Andressa Cerqueira† and Florencia Leonardi‡

Instituto de Matemática e Estatística, Universidade de São Paulo

May 22, 2018

Abstract

In this paper we introduce the Krichevsky-Trofimov estimator for the number of communities in the Stochastic Block Model (SBM) and prove its eventual almost sure convergence to the underlying number of communities, without assuming a known upper bound on that quantity. Our results apply to both the dense and the sparse regimes. To our knowledge this is the first strong consistency result for the estimation of the number of communities in the SBM, even in the bounded case.

Keywords: Stochastic Block Model, Krichevsky-Trofimov, Model selection.

1 Introduction

In this paper we address the model selection problem for the Stochastic Block Model (SBM); that is, the estimation of the number of communities given a sample of the adjacency matrix. The SBM was introduced by Holland *et al.* (1983) and has rapidly popularized in the literature as a model for random networks exhibiting blocks or communities between their nodes. In this model, each node in the network has associated a latent discrete random variable describing its community label, and given two nodes, the possibility of a connection between them depends only on the values of the nodes' latent variables.

From a statistical point of view, some methods have been proposed to address the problem of parameter estimation or label recovering for the SBM. Some examples include maximum likelihood estimation (Bickel & Chen, 2009; Amini *et al.* , 2013), variational methods (Daudin *et al.* , 2008; Latouche *et al.* , 2012), spectral clustering (Rohe *et al.*

*This work was produced as part of the activities of FAPESP Research, Innovation and Dissemination Center for Neuromathematics, grant 2013/07699-0, and FAPESP's project *Structure selection for stochastic processes in high dimensions*, grant 2016/17394-0.

†AC was supported by a FAPESP's scholarship (2015/12595-4).

‡FL was partially supported by a CNPq's research fellowship (309964/2016-4).

, 2011) and Bayesian inference (van der Pas *et al.* , 2017). The asymptotic properties of these estimators have also been considered in subsequent works such as Bickel *et al.* (2013) or Su *et al.* (2017). All these approaches assume the number of communities is known *a priori*.

The model selection problem for the SBM, that is the estimation of the number of communities, was also addressed before, see for example the recent work Le & Levina (2015) and references therein. But to our knowledge it was not until Wang *et al.* (2017) that a consistency result was obtained for such a penalized estimator. In the latter, the authors propose a penalized likelihood criterion and show its convergence in probability (weak consistency) to the true number of communities. Their proof only applies to the case where the number of candidate values for the estimator is finite (it is upper bounded by a known constant) and the network average degree grows at least as a polylog function on the number of nodes. From a practical point of view, the computation of the log-likelihood function and its supremum is not a simple task due to the hidden nature of the nodes' labels. Wang *et al.* (2017) propose a variational method as described in Bickel *et al.* (2013) using the EM algorithm of Daudin *et al.* (2008), a profile maximum likelihood criterion as in Bickel & Chen (2009) or the pseudo-likelihood algorithm in Amini *et al.* (2013). The method introduced in Wang *et al.* (2017) has been subsequently studied in Hu *et al.* (2016), where the authors propose a modification of the penalty term. However, in practice, the computation of the suggested estimator still remains a demanding task since it depends on the profile maximum likelihood function.

In this paper we take an information-theoretic perspective and introduce the Krichevsky-Trofimov (KT) estimator, see Krichevsky & Trofimov (1981), in order to determine the number of communities of a SBM based on a sample of the adjacency matrix of the network. We prove the strong consistency of this estimator, in the sense that the empirical value is equal to the correct number of communities in the model with probability one, as long as the number of nodes n in the network is sufficiently large. The strong consistency is proved in the *dense* regime, where the probability of having an edge is considered to be constant, and in the *sparse* regime where this probability goes to zero with n having order ρ_n. The study of the second regime is more interesting in the sense that it is necessary to control how much information is required (in the sense of the number of edges in the network) to estimate the parameters of the model. We prove that the consistency in the sparse case is guaranteed when the expected degree of a random selected node grows to infinity as a function of order $n\rho_n \to \infty$, weakening the assumption in Wang *et al.* (2017) that proves consistency in the regime $\frac{n\rho_n}{\log n} \to \infty$. We also consider a smaller order penalty function and we do not assume a known upper bound on the true number of communities. To our knowledge this is the first strong consistency result for an estimator of the number of communities, even in the bounded case.

The paper is organized as follows. In Section 2 we define the model and the notation used in the paper, in Section 3 we introduce the KT estimator for the number of communities and state the main result. The proof of the consistency of the estimator is presented in Section 4.

2 The Stochastic Block Model

Consider a non-oriented random network with nodes $\{1, 2, \ldots, n\}$, specified by its adjacency matrix $\mathbf{X}_{n\times n} \in \{0,1\}^{n\times n}$ that is symmetric and has diagonal entries equal to zero. Each node i has associated a latent (non-observed) variable Z_i on $[k] := \{1, 2, \ldots, k\}$, the *community* label of node i.

The SBM with k communities is a probability model for a random network as above, where the latent variables $\mathbf{Z}_n = (Z_1, Z_2, \cdots, Z_n)$ are independent and identically distributed random variables over $[k]$ and the law of the adjacency matrix $\mathbf{X}_{n\times n}$, conditioned on the value of the latent variables $\mathbf{Z}_n = \mathbf{z}_n$, is a product measure of Bernoulli random variables whose parameters depend only on the nodes' labels. More formally, there exists a probability distribution over $[k]$, denoted by $\pi = (\pi_1, \cdots, \pi_k)$, and a symmetric probability matrix $P \in [0,1]^{k\times k}$ such that the distribution of the pair $(\mathbf{Z}_n, \mathbf{X}_{n\times n})$ is given by

$$\mathbb{P}_{\pi,P}(\mathbf{z}_n, \mathbf{x}_{n\times n}) = \prod_{a=1}^{k} \pi_a^{n_a} \prod_{a,b=1}^{k} P_{a,b}^{O_{a,b}/2}(1-P_{a,b})^{(n_{a,b}-O_{a,b})/2}\,, \tag{2.1}$$

where the counters $n_a = n_a(\mathbf{z}_n)$, $n_{a,b} = n_{a,b}(\mathbf{z}_n)$ and $O_{a,b} = O_{a,b}(\mathbf{z}_n, \mathbf{x}_{n\times n})$ are given by

$$n_a(\mathbf{z}_n) = \sum_{i=1}^{n} \mathbb{1}\{z_i = a\}\,, \qquad 1 \le a \le k$$

$$n_{a,b}(\mathbf{z}_n) = \begin{cases} n_a(\mathbf{z}_n)n_b(\mathbf{z}_n)\,, & 1 \le a,b \le k\,;\, a \ne b \\ n_a(\mathbf{z}_n)(n_a(\mathbf{z}_n)-1) & 1 \le a,b \le k\,;\, a = b \end{cases}$$

and

$$O_{a,b}(\mathbf{z}_n, \mathbf{x}_{n\times n}) = \sum_{i,j=1}^{n} \mathbb{1}\{z_i = a, z_j = b\}x_{ij}\,, \quad 1 \le a,b \le k\,.$$

As it is usual in the definition of likelihood functions, by convention we define $0^0 = 1$ in (2.1) when some of the parameters are 0.

We denote by Θ^k the parametric space for a model with k communities, given by

$$\Theta^k = \left\{ (\pi, P)\colon \pi \in (0,1]^k,\, \sum_{a=1}^{k} \pi_a = 1,\, P \in [0,1]^{k\times k},\, P \text{ symmetric} \right\}.$$

The *order* of the SBM is defined as the smallest k for which the equality (2.1) holds for a pair of parameters $(\pi^0, P^0) \in \Theta^k$ and will be denoted by k_0. If a SBM has order k_0 then it cannot be reduced to a model with less communities than k_0; this specifically means that P^0 does not have two identical columns.

When P^0 is fixed and does not depend on n, the mean degree of a given node grows linearly in n and this regime produces very connected (dense graphs). For this reason in this paper we also consider the regime producing sparse graphs (with less edges), that is we allow P^0 to decrease with n to the zero matrix. In this case we write $P^0 = \rho_n S^0$, where $S^0 \in [0,1]^{k\times k}$ does not depend on n and ρ_n is a function decreasing to 0 at a rate $n\rho_n \to \infty$.

3 The KT order estimator

The Krichevsky-Trofimov order estimator in the context of a SBM is a regularized estimator based on a mixture distribution for the adjacency matrix $\mathbf{X}_{n\times n}$. Given a sample $(\mathbf{z}_n, \mathbf{x}_{n\times n})$ from the distribution (2.1) with parameters (π^0, P^0), where we assume we only observed the network $\mathbf{x}_{n\times n}$, the estimator of the number of communities is defined by

$$\hat{k}_{\mathrm{KT}}(\mathbf{x}_{n\times n}) = \arg\max_k \{ \log \mathrm{KT}_k(\mathbf{x}_{n\times n}) - \mathrm{pen}(k,n) \}\,, \tag{3.1}$$

where $\mathrm{KT}_k(\mathbf{x}_{n\times n})$ is the mixture distribution for a SBM with k communities and $\mathrm{pen}(k,n)$ is a penalizing function that will be specified later.

As it is usual for the KT distributions we choose as "prior" for the pair (π, P) a product measure obtained by a Dirichlet$(1/2,\cdots,1/2)$ distribution (the prior distribution for π) and a product of $(k^2+k)/2$ Beta$(1/2,1/2)$ distributions (the prior for the symmetric matrix P). In other words, we define the distribution on Θ^k

$$\nu_k(\pi,P) = \left[\frac{\Gamma(\frac{k}{2})}{\Gamma(\frac{1}{2})^k}\prod_{a=1}^{k}\pi_a^{-\frac{1}{2}}\right]\left[\prod_{1\le a\le b\le k}\frac{1}{\Gamma(\frac{1}{2})^2}\,P_{a,b}^{-\frac{1}{2}}(1-P_{a,b})^{-\frac{1}{2}}\right] \tag{3.2}$$

and we construct the mixture distribution for $\mathbf{X}_{n\times n}$, based on $\nu_k(\pi,P)$, given by

$$\mathrm{KT}_k(\mathbf{x}_{n\times n}) = \mathbb{E}_{\nu_k}[\,\mathbb{P}_{\pi,P}(\mathbf{x}_{n\times n})\,] = \int_{\Theta^k}\mathbb{P}_{\pi,P}(\mathbf{x}_{n\times n})\nu_k(\pi,P)d\pi dP\,, \tag{3.3}$$

where $\mathbb{P}_{\pi,P}(\mathbf{x}_{n\times n})$ stands for the marginal distribution obtained from (2.1), and given by

$$\mathbb{P}_{\pi,P}(\mathbf{x}_{n\times n}) = \sum_{\mathbf{z}_n\in[k]^n}\mathbb{P}_{\pi,P}(\mathbf{z}_n,\mathbf{x}_{n\times n})\,. \tag{3.4}$$

As in other model selection problems where the KT approach has proved to be very useful, as for example in the case of Context Tree Models (Csiszar & Talata, 2006) or Hidden Markov models (Gassiat & Boucheron, 2003), in the case of the SBM there is a closed relationship between the KT mixture distribution and the maximum likelihood function. The following proposition shows a non asymptotic uniform upper bound for the log ratio between these two functions. Its proof is postponed to the Appendix.

Proposition 3.1. *For all k and all $n\ge \max(4,k)$ we have*

$$\max_{\mathbf{x}_{n\times n}}\left\{\log\frac{\sup_{(\pi,P)\in\Theta^k}\mathbb{P}_{\pi,P}(\mathbf{x}_{n\times n})}{\mathrm{KT}_k(\mathbf{x}_{n\times n})}\right\} \le \left(\tfrac{k(k+2)}{2}-\tfrac{1}{2}\right)\log n + c_{k,n}\,,$$

where

$$c_{k,n} = \tfrac{k(k+1)}{2}\log\Gamma\left(\tfrac{1}{2}\right) + \tfrac{k(k-1)}{4n} + \tfrac{1}{12n} + \log\frac{\Gamma(\frac{1}{2})}{\Gamma(\frac{k}{2})} + \tfrac{7k(k+1)}{12}\,.$$

Proposition 3.1 is at the core of the proof of the strong consistency of $\hat{k}_{\mathrm{KT}}$ defined by (3.1). By strong consistency we mean that the estimator equals the order k_0 of the SBM with probability one, for all sufficiently large n (that may depend on the sample $\mathbf{x}_{n\times n}$). In order to derive the strong consistency result for the KT order estimator, we need a penalty function in (3.1) with a given rate of convergence when n grows to infinity. Although there are a range of possibilities for this penalty function, the specific form we use in this paper is

$$\begin{aligned} \mathrm{pen}(k,n) &= \sum_{i=1}^{k-1} \frac{(i(i+2)+3+\epsilon)}{2} \log n \\ &= \left[\frac{k(k-1)(2k-1)}{12} + \frac{k(k-1)}{2} + \frac{(3+\epsilon)(k-1)}{2}\right] \log n \end{aligned} \tag{3.5}$$

for some $\epsilon > 0$. The convenience of the expression above will be make clear in the proof of the consistency result. Observe that the penalty function defined by (3.5) is dominated by a tern of order $k^3 \log n$ and then it is of smaller order than the function $\frac{k(k+1)}{2} n \log n$ used in Wang *et al.* (2017), so our results also apply in this case. It remains an open question which is the smallest penalty function for a strongly consistent estimator.

We finish this section by stating the main theoretical result in this paper.

Theorem 3.2 (Consistency Theorem). *Suppose the SBM has order k_0 with parameters (π^0, P^0). Then, for a penalty function of the form (3.5) we have that*

$$\hat{k}_{KT}(\mathbf{x}_{n\times n}) = k_0$$

eventually almost surely as $n \to \infty$.

The proof of this and other auxiliary results are given in the next section and in the Appendix.

4 Proof of the Consistency Theorem

The proof of Theorem 3.2 is divided in two main parts. The first one, presented in Subsection 4.1, proves that $\hat{k}_{\mathrm{KT}}(\mathbf{x}_{n\times n})$ does not overestimate the true order k_0, eventually almost surely when $n \to \infty$, even without assuming a known upper bound on k_0. The second part of the proof, presented in Subsection 4.2, shows that $\hat{k}_{\mathrm{KT}}(\mathbf{x}_{n\times n})$ does not underestimate k_0, eventually almost surely when $n \to \infty$. By combining these two results we prove that $\hat{k}_{\mathrm{KT}}(\mathbf{x}_{n\times n}) = k_0$ eventually almost surely as $n \to \infty$.

4.1 Non-overestimation

The main result in this subsection is given by the following proposition.

Proposition 4.1. *Let $\mathbf{x}_{n\times n}$ be a sample of size n from a SBM of order k_0, with parameters π^0 and P^0. Then, the $\hat{k}_{KT}(\mathbf{x}_{n\times n})$ order estimator defined in (3.1) with penalty function given by (3.5) does not overestimate k_0, eventually almost surely when $n \to \infty$.*

The proof of Proposition 4.1 follows straightforward from Lemmas 4.2, 4.3 and 4.4 presented below. These lemmas are inspired in the work Gassiat & Boucheron (2003) which proves consistency for an order estimator of a Hidden Markov Model.

Lemma 4.2. *Under the hypotheses of Proposition 4.1 we have that*

$$\hat{k}_{KT}(\mathbf{x}_{n\times n}) \notin (k_0, \log n]$$

eventually almost surely when $n \to \infty$.

Proof. First observe that

$$\mathbb{P}_{\pi^0,P^0}(\hat{k}_{\mathrm{KT}}(\mathbf{x}_{n\times n}) \in (k_0, \log n]) \;=\; \sum_{k=k_0+1}^{\log n} \mathbb{P}_{\pi^0,P^0}(\hat{k}_{\mathrm{KT}}(\mathbf{x}_{n\times n}) = k)\,. \tag{4.1}$$

Using Lemma A.2 we can bound the sum in the right-hand side by

$$\begin{aligned}\sum_{k=k_0+1}^{\log n} &\exp\Big\{\tfrac{(k_0(k_0+2)-1)}{2}\log n + c_{k_0,n} + \mathrm{pen}(k_0,n) - \mathrm{pen}(k,n)\Big\}\\ &\le e^{c_{k_0,n}}\,\log n\,\exp\Big\{\tfrac{(k_0(k_0+2)-1)}{2}\log n + \mathrm{pen}(k_0,n) - \mathrm{pen}(k_0+1,n)\Big\}\end{aligned}$$

where the last inequality follows from the fact that $\mathrm{pen}(k,n)$ is an increasing function in k. Moreover, a simple calculation using the specific form in (3.5) gives

$$\begin{aligned}\tfrac{(k_0(k_0+2)-1)}{2}&\log n + \mathrm{pen}(k_0,n) - \mathrm{pen}(k_0+1,n)\\ &= \big(\tfrac{(k_0(k_0+2)-1)}{2} - \tfrac{(k_0(k_0+2)-1+4+\epsilon))}{2}\big)\log n\\ &= -(2+\epsilon/2)\log n\,.\end{aligned}$$

By using this expression in the right-hand side of the las inequality to bound (4.1) we obtain that

$$\sum_{n=1}^{\infty} \mathbb{P}_{\pi^0,P^0}(\hat{k}_{\mathrm{KT}}(\mathbf{x}_{n\times n}) \in (k_0, \log n]) \;\le\; C_{k_0}\sum_{n=1}^{\infty}\frac{\log n}{n^{2+\epsilon/2}} \;<\; \infty\,,$$

where C_{k_0} denotes an upper-bound on $\exp(c_{k_0,n})$. Now the result follows by the first Borel Cantelli lemma. □

Lemma 4.3. *Under the hypotheses of Proposition 4.1 we have that*

$$\hat{k}_{KT}(\mathbf{x}_{n\times n}) \notin (\log n, n]$$

eventually almost surely when $n \to \infty$.

Proof. As in the proof of Lemma 4.2 we write

$$\mathbb{P}_{\pi^0,P^0}(\hat{k}_{\mathrm{KT}}(\mathbf{x}_{n\times n}) \in (\log n, n]) = \sum_{k=\log n}^{n} \mathbb{P}_{\pi^0,P^0}(\hat{k}_{\mathrm{KT}}(\mathbf{x}_{n\times n}) = k)$$

and we use again Lemma A.2 to bound the sum in the right-hand side by

$$\begin{aligned}&\sum_{k=\log n}^{n} \exp\Big\{\tfrac{(k_0(k_0+2)-1)}{2}\log n + c_{k_0,n} + \mathrm{pen}(k_0,n) - \mathrm{pen}(k,n)\Big\}\\ &\qquad \le\; e^{c_{k_0,n}}\, n\, \exp\Big\{-\log n\Big[-\tfrac{(k_0(k_0+2)-1)}{2} - \tfrac{\mathrm{pen}(k_0,n)}{\log n} + \tfrac{\mathrm{pen}(\log n,n)}{\log n}\Big]\Big\}\,.\end{aligned}$$

Since $\mathrm{pen}(k,n)/\log(n)$ does not depend on n and increases cubically in k we have that

$$\liminf_{n\to\infty} \tfrac{\mathrm{pen}(\log n,n)}{\log n} - \tfrac{(k_0(k_0+2)-1)}{2} - \tfrac{\mathrm{pen}(k_0,n)}{\log n} \;>\; 3$$

and thus

$$\sum_{n=1}^{\infty} n\, \exp\Big\{-\log n\Big[-\tfrac{(k_0(k_0+2)-1)}{2} - \tfrac{\mathrm{pen}(k_0,n)}{\log n} + \tfrac{\mathrm{pen}(\log n,n)}{\log n}\Big]\Big\} \;<\; \infty\,.$$

Using the fact that $\exp(c_{k_0,n})$ is decreasing on n, the result follows from the first Borel Cantelli lemma. □

Lemma 4.4. *Under the hypotheses of Proposition 4.1 we have that*

$$\hat{k}_{KT}(\mathbf{x}_{n\times n}) \notin (n,\infty)$$

eventually almost surely when $n\to\infty$.

Proof. Observe that it is enough to prove that

$$\log \mathrm{KT}_{n+m}(\mathbf{x}_{n\times n}) - \mathrm{pen}(n+m,n) \;\le\; \log \mathrm{KT}_n(\mathbf{x}_{n\times n}) - \mathrm{pen}(n,n)$$

for all $m\ge 1$. By using Proposition 3.1 we have that

$$-\log \mathrm{KT}_n(\mathbf{x}_{n\times n}) \;\le\; -\log \sup_{(\pi,P)\in\Theta^n} \mathbb{P}_{\pi,P}(\mathbf{x}_{n\times n}) + \Big(\tfrac{n(n+2)}{2} - \tfrac{1}{2}\Big)\log n + c_{n,n}$$

and by (3.3) we obtain

$$\mathrm{KT}_{n+m}(\mathbf{x}_{n\times n}) \;\le\; \sup_{(\pi,P)\in\Theta^{n+m}} \mathbb{P}_{\pi,P}(\mathbf{x}_{n\times n})\,.$$

Thus, as

$$\sup_{(\pi,P)\in\Theta^{n+m}} \mathbb{P}_{\pi,P}(\mathbf{x}_{n\times n}) \;=\; \sup_{(\pi,P)\in\Theta^{n}} \mathbb{P}_{\pi,P}(\mathbf{x}_{n\times n})$$

we obtain

$$\begin{aligned}\log\ \mathrm{KT}_{n+m}(\mathbf{x}_{n\times n}) - \log \mathrm{KT}_n(\mathbf{x}_{n\times n}) \;&\le\; \Big(\tfrac{n(n+2)}{2} - \tfrac{1}{2}\Big)\log n + c_{n,n}\\ \le\; \tfrac{(n(n+2)-1)}{2}\log n + n(n+1)\Big(\tfrac{\log\Gamma(\frac{1}{2})}{2} + \tfrac{7}{12}\Big) &+ \tfrac{n(n-1)}{4n} + \tfrac{1}{12n} - \log\tfrac{\Gamma(\frac{n}{2})}{\Gamma(\frac{1}{2})}\\ \le\; \mathrm{pen}(n+m,n) - \mathrm{pen}(n,n)&\end{aligned}$$

where the last inequality holds for n big enough. □

4.2 Non-underestimation

In this subsection we deal with the proof of the non-underestimation of $\hat{k}_{\text{KT}}(\mathbf{x}_{n\times n})$. The main result of this section is the following

Proposition 4.5. *Let $\mathbf{x}_{n\times n}$ be a sample of size n from a SBM of order k_0 with parameters (π^0, P^0). Then, the $\hat{k}_{KT}(\mathbf{x}_{n\times n})$ order estimator defined in (3.1) with penalty function given by (3.5) does not underestimate k_0, eventually almost surely when $n \to \infty$.*

In order to prove this result we need Lemmas 4.6 and 4.7 below, that explore limiting properties of the under-fitted model. That is we handle with the problem of fitting a SBM of order k_0 in the parameter space Θ^{k_0-1}.

An intuitive construction of a $(k-1)$-block model from a k-block model is obtained by merging two given blocks. This merging can be implemented in several ways, but here we consider the construction given in Wang *et al.* (2017), with the difference that instead of using the sample block proportions we use the limiting distribution π of the original k-block model.

Given $(\pi, P) \in \Theta^k$ we define the merging operation $M_{a,b}(\pi, P) = (\pi^*, P^*) \in \Theta^{k-1}$ which combines blocks with labels a and b. For ease of exposition we only show the explicit definition for the case $a = k-1$ and $b = k$. In this case, the merged distribution π^* is given by

$$\begin{aligned} \pi^*_i &= \pi_i && \text{for } 1 \le i \le k-2\,, \\ \pi^*_{k-1} &= \pi_{k-1} + \pi_k\,. \end{aligned} \tag{4.2}$$

On the other hand, the merged matrix P^* is obtained as

$$\begin{aligned} P^*_{l,r} &= P_{l,r} && \text{for } 1 \le l, r \le k-2\,, \\ P^*_{l,k-1} &= \frac{\pi_l \pi_{k-1} P_{l,k-1} + \pi_l \pi_k P_{l,k}}{\pi_l \pi_{k-1} + \pi_l \pi_k} && \text{for } 1 \le l \le k-2\,, \\ P^*_{k-1,k-1} &= \frac{\pi_{k-1}\pi_{k-1} P_{k-1,k-1} + 2\pi_{k-1}\pi_k P_{k-1,k} + \pi_k \pi_k P_{k,k}}{\pi_{k-1}\pi_{k-1} + 2\pi_{k-1}\pi_k + \pi_k\pi_k}\,. \end{aligned} \tag{4.3}$$

For arbitrary a and b the definition is obtained by permuting the labels.

Given $\mathbf{x}_{n\times n}$ originated from the SBM of order k_0 and parameters (π^0, P^0), we define the profile likelihood estimator of the label assignment under the (k_0-1)-block model as

$$\mathbf{z}^\star_n = \operatorname*{arg\,max}_{\mathbf{z}_n \in [k_0-1]^n} \sup_{(\pi,P)\in\Theta^{k_0-1}} \mathbb{P}_{\pi,P}(\mathbf{z}_n, \mathbf{x}_{n\times n})\,. \tag{4.4}$$

The next lemmas show that the logarithm of the ratio between the maximum likelihood under the true order k_0 and the maximum profile likelihood under the under-fitting k_0-1 order model is bounded from below by a function growing faster than $n \log n$, eventually almost surely when $n \to \infty$. Each lemma consider one of the two possible regimes $\rho_n = \rho > 0$ (dense regime) or $\rho_n \to 0$ at a rate $n\rho_n \to \infty$ (sparse regime).

Lemma 4.6 (dense regime)**.** *Let $(\mathbf{z}_n, \mathbf{x}_{n\times n})$ be a sample of size n from a SBM of order k_0 with parameters (π^0, P^0), with P^0 not depending on n. Then there exist $r, s \in [k_0]$ such that for $(\pi^*, P^*) = M_{r,s}(\pi^0, P^0)$ we have that almost surely*

$$\begin{aligned}\liminf_{n\to\infty} \frac{1}{n^2} \log \frac{\sup_{(\pi,P)\in\Theta^{k_0}} \mathbb{P}_{\pi,P}(\mathbf{z}_n, \mathbf{x}_{n\times n})}{\sup_{(\pi,P)\in\Theta^{k_0-1}} \mathbb{P}_{\pi,P}(\mathbf{z}^\star_n, \mathbf{x}_{n\times n})} & \\ \geq \frac{1}{2}\Bigg[\sum_{a,b=1}^{k_0} \pi^0_a\pi^0_b\,\gamma(P^0_{ab}) - \sum_{a,b=1}^{k_0-1} \pi^*_a\pi^*_b\,\gamma(P^*_{a,b})\Bigg] & \qquad (4.5)\\ > 0\,, &\end{aligned}$$

where $\gamma(x) = x\log x + (1-x)\log(1-x)$.

Proof. Given k and $\bar{\mathbf{z}}_n \in [k]^n$ define the empirical probabilities

$$\begin{aligned}\hat{\pi}_a(\bar{\mathbf{z}}_n) &= \frac{n_a(\bar{\mathbf{z}}_n)}{n}\,, && 1 \leq a \leq k\\ \hat{P}_{a,b}(\bar{\mathbf{z}}_n, \mathbf{x}_{n\times n}) &= \frac{O_{a,b}(\bar{\mathbf{z}}_n, \mathbf{x}_{n\times n})}{n_{a,b}(\bar{\mathbf{z}}_n)}\,, && 1 \leq a, b \leq k\,.\end{aligned} \qquad (4.6)$$

Then the maximum likelihood function is given by

$$\begin{aligned}\log \sup_{(\pi,P)\in\Theta^{k_0}} \mathbb{P}_{\pi,P}(\mathbf{z}_n, \mathbf{x}_{n\times n}) \;=\; & n\sum_{a=1}^{k_0} \hat{\pi}_a(\mathbf{z}_n)\log\hat{\pi}_a(\mathbf{z}_n)\\ & + \frac{1}{2}\sum_{a,b=1}^{k_0} n_{a,b}(\mathbf{z}_n)\gamma(\hat{P}_{a,b}(\mathbf{z}_n, \mathbf{x}_{n\times n}))\,.\end{aligned}$$

Using that $n_{a,b} = n_a n_b$ for $a \neq b$ and $n_{a,a} = n_a(n_a - 1)$ the last expression is equal to

$$\begin{aligned} n\sum_{a=1}^{k_0} \hat{\pi}_a(\mathbf{z}_n)\log\hat{\pi}_a(\mathbf{z}_n) - \frac{n}{2}\sum_{a=1}^{k_0} \hat{\pi}_a(\mathbf{z}_n)\gamma(\hat{P}_{a,a}(\mathbf{z}_n, \mathbf{x}_{n\times n}))\\ + \frac{n^2}{2}\sum_{a,b=1}^{k_0} \hat{\pi}_a(\mathbf{z}_n)\hat{\pi}_b(\mathbf{z}_n)\gamma(\hat{P}_{a,b}(\mathbf{z}_n, \mathbf{x}_{n\times n}))\,.\end{aligned} \qquad (4.7)$$

The first two terms in (4.7) are of smaller order compared to n^2, so by the Strong Law of Large Numbers we have that almost surely

$$\lim_{n\to\infty} \frac{1}{n^2} \log \sup_{(\pi,P)\in\Theta^{k_0}} \mathbb{P}_{\pi,P}(\mathbf{z}_n, \mathbf{x}_{n\times n}) \;=\; \frac{1}{2}\sum_{a,b=1}^{k_0} \pi^0_a\pi^0_b\,\gamma(P^0_{a,b})\,. \qquad (4.8)$$

Similarly for $k_0 - 1$ and $\mathbf{z}^\star_n \in [k_0-1]^n$ we have that almost surely

$$\limsup_{n\to\infty} \frac{1}{n^2} \log \sup_{(\pi,P)\in\Theta^{k_0-1}} \mathbb{P}_{\pi,P}(\mathbf{z}^\star_n, \mathbf{x}_{n\times n}) = \frac{1}{2}\sum_{a,b=1}^{k_0-1} \tilde{\pi}_a\tilde{\pi}_b\,\gamma(\tilde{P}_{a,b})\,, \qquad (4.9)$$

for some $(\tilde\pi, \tilde P) \in \Theta^{k_0-1}$. Combining (4.8) and (4.9) we have that almost surely

$$
\begin{aligned}
\liminf_{n\to\infty} \frac{1}{n^2} \log \frac{\sup_{(\pi,P)\in\Theta^{k_0}} \mathbb{P}_{\pi,P}(\mathbf{z}_n, \mathbf{x}_{n\times n})}{\sup_{(\pi,P)\in\Theta^{k_0-1}} \mathbb{P}_{\pi,P}(\mathbf{z}_n^\star, \mathbf{x}_{n\times n})} \\
= \frac{1}{2} \sum_{a,b=1}^{k_0} \pi_a^0 \pi_b^0 \, \gamma(P_{a,b}^0) - \frac{1}{2} \sum_{a,b=1}^{k_0-1} \tilde\pi_a \tilde\pi_b \, \gamma(\tilde P_{a,b}) \,.
\end{aligned}
\tag{4.10}
$$

To obtain a lower bound for (4.10) we need to compute $(\tilde\pi, \tilde P)$ that minimizes the right-hand side. This is equivalent to obtain $(\tilde\pi, \tilde P) \in \Theta^{k_0-1}$ that maximizes the second term

$$
\sum_{a,b=1}^{k_0-1} \tilde\pi_a \tilde\pi_b \, \gamma(\tilde P_{a,b}) \,. \tag{4.11}
$$

Denote by $(\widetilde{\mathbf{Z}}_n, \widetilde{\mathbf{X}}_{n\times n})$ a (k_0-1)-order SBM with distribution $(\tilde\pi, \tilde P)$. By definition

$$
\tilde P_{\tilde a, \tilde b} = \frac{P(\tilde X_{i,j} = 1, \tilde Z_i = \tilde a, \tilde Z_j = \tilde b)}{P(\tilde Z_i = \tilde a, \tilde Z_j = \tilde b)} \,.
$$

Observe that when $\widetilde{\mathbf{X}}_{n\times n} = \mathbf{X}_{n\times n}$, the numerator equals

$$
\begin{aligned}
&\sum_{a,b=1}^{k_0} P(X_{i,j} = 1 | Z_i = a, Z_j = b) P(Z_i = a, Z_j = b, \tilde Z_i = \tilde a, \tilde Z_j = \tilde b) \\
&\quad = \sum_{a,b=1}^{k_0} P(Z_i = a, \tilde Z_i = \tilde a) \, P_{a,b}^0 \, P(Z_j = b, \tilde Z_j = \tilde b) \\
&\quad = (Q P^0 Q^T)_{\tilde a, \tilde b} \,,
\end{aligned}
$$

where $Q(a, \tilde a)$ denotes a joint distribution on $[k_0] \times [k_0 - 1]$ (a coupling) with marginals π^0 and $\tilde\pi$, respectively. Similarly, the denominator can be written as

$$
\sum_{a,b=1}^{k_0} P(Z_i = a, \tilde Z_i = \tilde a) P(Z_j = b, \tilde Z_j = \tilde b) = (Q(\mathbf{1}\mathbf{1}^T) Q^T)_{\tilde a, \tilde b} \,,
$$

where $\mathbf{1}$ denotes the matrix with dimension $(k_0 - 1) \times k_0$ and all entries equal to 1. Then we can rewrite (4.11) as

$$
\sum_{a,b=1}^{k_0-1} (Q(\mathbf{1}\mathbf{1}^T)Q^T)_{a,b} \, \gamma \left[\frac{(QP^0Q^T)_{a,b}}{(Q(\mathbf{1}\mathbf{1}^T)Q^T)_{a,b}} \right] . \tag{4.12}
$$

Therefore, finding a pair $(\tilde\pi, \tilde P)$ maximizing (4.11) is equivalent to finding an optimal coupling Q maximizing (4.12). Wang *et al.* (2017) proved that there exist $r, s \in [k_0]$

such that (4.12) achieves its maximum at $(\pi^*, P^*) = M_{r,s}(\pi^0, P^0)$, see Lemma A.2 there. This concludes the proof of the first inequality in (4.5). In order to prove the second strict inequality in (4.5), we consider for convenience and without loss of generality, $r = k_0 - 1$ and $s = k_0$ (the other cases can be handled by a permutation of the labels). Notice that in the right-hand side of (4.10), with $(\tilde{\pi}, \tilde{P})$ substituted by the optimal value $M_{k_0-1,k_0}(\pi^0, P^0)$ defined by (4.2) and (4.3), all the terms with $1 \le a, b \le k_0 - 2$ cancel. Moreover, as γ is a convex function, Jensen's inequality implies that

$$\pi^*_a \pi^*_{k_0-1} \gamma(P^*_{a,k_0-1}) \;\le\; \pi^0_a \pi^0_{k_0-1} \gamma(P^0_{a,k_0-1}) + \pi^0_a \pi^0_{k_0} \gamma(P^0_{a,k_0}) \tag{4.13}$$

for all $a = 1, \ldots, k_0 - 2$ and similarly

$$(\pi^*_{k_0-1})^2 \gamma(P^*_{k_0-1,k_0-1}) \;\le\; \sum_{a,b=k_0-1}^{k_0} \pi^0_a \pi^0_b \gamma(P^0_{a,b})\,. \tag{4.14}$$

The equality holds for all a in (4.13) and in (4.14) simultaneously if and only if

$$P^0_{a,k_0} \;=\; P_{a,k_0-1} \qquad \text{for all } a = 1, \ldots, k_0\,,$$

in which case the matrix P^0 would have two identical columns, contradicting the fact that the sample $(\mathbf{z}_n, \mathbf{x}_{n\times n})$ originated from a SBM with order k_0. Therefore the strict inequality must hold in (4.13) for at least one a or in (4.14), showing that the second inequality in (4.5) holds. □

Lemma 4.7 (sparse regime)**.** *Let $(\mathbf{z}_n, \mathbf{x}_{n\times n})$ be a sample of size n from a SBM of order k_0 with parameters $(\pi^0, \rho_n S^0)$, where $\rho_n \to 0$ at a rate $n\rho_n \to \infty$. Then there exist $r, s \in [k_0]$ such that for $(\pi^*, P^*) = M_{r,s}(\pi^0, S^0)$ we have that almost surely*

$$\begin{aligned}\liminf_{n\to\infty} \frac{1}{\rho_n n^2} \log \frac{\sup_{(\pi,P)\in\Theta^{k_0}} \mathbb{P}_{\pi,P}(\mathbf{z}_n, \mathbf{x}_{n\times n})}{\sup_{(\pi,P)\in\Theta^{k_0-1}} \mathbb{P}_{\pi,P}(\mathbf{z}^\star_n, \mathbf{x}_{n\times n})} &\\ \ge\; \frac{1}{2}\left[\sum_{a,b=1}^{k_0} \pi^0_a \pi^0_b\, \tau(S^0_{a,b}) - \sum_{a,b=1}^{k_0-1} \pi^*_a \pi^*_b\, \tau(P^*_{a,b})\right] &\\ >\; 0\,, &\end{aligned} \tag{4.15}$$

where $\tau(x) = x \log x - x$.

Proof. This proof follows the same arguments used in the proof of Lemma 4.6, but as in this case P^0 decreases to 0 some limits must be handled differently. As shown in (4.7) we have that

$$\log \sup_{(\pi,P)\in\Theta^{k_0}} \mathbb{P}_{\pi,P}(\mathbf{z}_n, \mathbf{x}_{n\times n}) = n \sum_{a=1}^{k_0} \hat{\pi}_a(\mathbf{z}_n) \log \hat{\pi}_a(\mathbf{z}_n) - \frac{n}{2} \sum_{a=1}^{k_0} \hat{\pi}_a(\mathbf{z}_n) \gamma(\hat{P}_{a,a}(\mathbf{z}_n, \mathbf{x}_{n\times n})) + \frac{n^2}{2} \sum_{a,b=1}^{k_0} \hat{\pi}_a(\mathbf{z}_n)\hat{\pi}_b(\mathbf{z}_n)\gamma(\hat{P}_{a,b}(\mathbf{z}_n, \mathbf{x}_{n\times n}))\,. \tag{4.16}$$

For $\rho_n \to 0$, Bickel & Chen (2009) proved that

$$\sum_{a,b=1}^{k_0} \hat{\pi}_a(\mathbf{z}_n)\hat{\pi}_b(\mathbf{z}_n)\gamma(\hat{P}_{a,b}(\mathbf{z}_n, \mathbf{x}_{n\times n})) = \rho_n \sum_{a,b=1}^{k_0} \hat{\pi}_a(\mathbf{z}_n)\hat{\pi}_b(\mathbf{z}_n)\,\tau\Big(\frac{\hat{P}_{a,b}(\mathbf{z}_n, \mathbf{x}_{n\times n})}{\rho_n}\Big) + \frac{E_n}{n^2}\log\rho_n + O(\rho_n^2)\,, \tag{4.17}$$

where $E_n = \sum_{a,b=1}^{k_0} O_{ab}(\mathbf{z}_n, \mathbf{x}_{n\times n})$ (twice the total number of edges in the graph) and $\tau(x) = x\log x - x$. Thus, as $\rho_n n \to \infty$ we can drop the first two terms in (4.16) and we have that

$$\frac{1}{\rho_n n^2} \log \sup_{(\pi,P)\in\Theta^{k_0}} \mathbb{P}_{\pi,P}(\mathbf{z}_n, \mathbf{x}_{n\times n}) = \frac{1}{2} \sum_{a,b=1}^{k_0} \hat{\pi}_a(\mathbf{z}_n)\hat{\pi}_b(\mathbf{z}_n)\tau\Big(\frac{\hat{P}_{a,b}(\mathbf{z}_n, \mathbf{x}_{n\times n})}{\rho_n}\Big) + \frac{E_n\log\rho_n}{2\rho_n n^2} + O(\rho_n) \tag{4.18}$$

and

$$\frac{1}{\rho_n n^2} \log \sup_{(\pi,P)\in\Theta^{k_0-1}} \mathbb{P}_{\pi,P}(\mathbf{z}^\star_n, \mathbf{x}_{n\times n}) = \frac{1}{2} \sum_{a,b=1}^{k_0-1} \hat{\pi}_a(\mathbf{z}^\star_n)\hat{\pi}_b(\mathbf{z}^\star_n)\tau\Big(\frac{\hat{P}_{a,b}(\mathbf{z}^\star_n, \mathbf{x}_{n\times n})}{\rho_n}\Big) + \frac{E_n\log\rho_n}{2\rho_n n^2} + O(\rho_n)\,. \tag{4.19}$$

Now, as in the proof of Lemma 4.6 there must exist some $(\tilde{\pi}, \tilde{S}) \in \Theta^{k_0-1}$ such that almost surely

$$\liminf_{n\to\infty} \frac{1}{\rho_n n^2} \log \frac{\sup_{(\pi,P)\in\Theta^{k_0}} \mathbb{P}_{\pi,P}(\mathbf{z}_n, \mathbf{x}_{n\times n})}{\sup_{(\pi,P)\in\Theta^{k_0-1}} \mathbb{P}_{\pi,P}(\mathbf{z}^\star_n, \mathbf{x}_{n\times n})} = \frac{1}{2}\sum_{a,b=1}^{k_0} \pi^0_a\pi^0_b\,\tau(S^0_{a,b}) - \frac{1}{2}\sum_{a,b=1}^{k_0-1} \tilde{\pi}_a\tilde{\pi}_b\,\tau(\tilde{S}_{a,b})\,. \tag{4.20}$$

As before, we want to obtain $(\tilde{\pi}, \tilde{S}) \in \Theta^{k_0-1}$ that maximizes the second term in the right-hand side of the equality above. The rest of the proof here is analogous to that of Lemma 4.6, by observing that τ is also a convex function and therefore the lower bound on 0 in (4.15) also holds. □

Proof of Prosposition 4.5. To prove that $\hat{k}_{\mathrm{KT}}(\mathbf{x}_{n\times n})$ does not underestimate k_0 it is enough to show that for all $k < k_0$

$$\log \mathrm{KT}_{k_0}(\mathbf{x}_{n\times n}) - \mathrm{pen}(k_0, n) \;>\; \log \mathrm{KT}_{k}(\mathbf{x}_{n\times n}) - \mathrm{pen}(k, n)$$

eventually almost surely when $n \to \infty$. As

$$\lim_{n\to\infty} \frac{1}{\rho_n n^2}\Big[\mathrm{pen}(k_0, n) - \mathrm{pen}(k, n)\Big] \;=\; 0$$

this is equivalent to show that

$$\liminf_{n\to\infty} \frac{1}{\rho_n n^2} \log \frac{\mathrm{KT}_{k_0}(\mathbf{x}_{n\times n})}{\mathrm{KT}_{k}(\mathbf{x}_{n\times n})} \;>\; 0\,.$$

First note that the logarithm above can be written as

$$\log \frac{\mathrm{KT}_{k_0}(\mathbf{x}_{n\times n})}{\mathrm{KT}_{k}(\mathbf{x}_{n\times n})} \;=\; \log \frac{\mathrm{KT}_{k_0}(\mathbf{x}_{n\times n})}{\sup_{(\pi,P)\in\Theta^{k_0}} \mathbb{P}_{\pi,P}(\mathbf{x}_{n\times n})} + \log \frac{\sup_{(\pi,P)\in\Theta^{k_0}} \mathbb{P}_{\pi,P}(\mathbf{x}_{n\times n})}{\mathrm{KT}_{k}(\mathbf{x}_{n\times n})}\,.$$

Using Proposition (3.1) we have that the first term in the right-hand side can be bounded below by

$$\log \frac{\mathrm{KT}_{k_0}(\mathbf{x}_{n\times n})}{\sup_{(\pi,P)\in\Theta^{k_0}} \mathbb{P}_{\pi,P}(\mathbf{x}_{n\times n})} \;\geq\; -\Big(\tfrac{k_0(k_0+2)}{2} - \tfrac{1}{2}\Big)\log n - c_{k_0,n}\,. \tag{4.21}$$

On the other hand, the second term can be lower-bounded by using

$$\begin{aligned}
\log & \frac{\sup_{(\pi,P)\in\Theta^{k_0}} \mathbb{P}_{\pi,P}(\mathbf{x}_{n\times n})}{\mathrm{KT}_{k}(\mathbf{x}_{n\times n})} \\
&= \log \frac{\sup_{(\pi,P)\in\Theta^{k_0}} \mathbb{P}_{\pi,P}(\mathbf{x}_{n\times n})}{\sup_{(\pi,P)\in\Theta^{k}} \mathbb{P}_{\pi,P}(\mathbf{x}_{n\times n})} + \log \frac{\sup_{(\pi,P)\in\Theta^{k}} \mathbb{P}_{\pi,P}(\mathbf{x}_{n\times n})}{\mathrm{KT}_{k}(\mathbf{x}_{n\times n})} \\
&\geq \log \frac{\sup_{(\pi,P)\in\Theta^{k_0}} \mathbb{P}_{\pi,P}(\mathbf{x}_{n\times n})}{\sup_{(\pi,P)\in\Theta^{k}} \mathbb{P}_{\pi,P}(\mathbf{x}_{n\times n})}\,.
\end{aligned} \tag{4.22}$$

By combining (4.21) and (4.22) we obtain

$$\frac{1}{\rho_n n^2} \log \frac{\mathrm{KT}_{k_0}(\mathbf{x}_{n\times n})}{\mathrm{KT}_{k'}(\mathbf{x}_{n\times n})} \;\geq\; -\Big(\tfrac{k_0(k_0+2)}{2} - \tfrac{1}{2}\Big)\frac{\log n}{\rho_n n^2} - \frac{c_{k_0,n}}{\rho_n n^2} + \frac{1}{\rho_n n^2} \log \frac{\sup_{(\pi,P)\in\Theta^{k_0}} \mathbb{P}_{\pi,P}(\mathbf{x}_{n\times n})}{\sup_{(\pi,P)\in\Theta^{k}} \mathbb{P}_{\pi,P}(\mathbf{x}_{n\times n})}\,.$$

Now, as $n\rho_n \to \infty$ it suffices to show that for $k < k_0$, almost surely we have

$$\liminf_{n\to\infty} \frac{1}{\rho_n n^2} \log \frac{\sup_{(\pi,P)\in\Theta^{k_0}} \mathbb{P}_{\pi,P}(\mathbf{x}_{n\times n})}{\sup_{(\pi,P)\in\Theta^{k}} \mathbb{P}_{\pi,P}(\mathbf{x}_{n\times n})} \; > \; 0\,. \tag{4.23}$$

We start with $k = k_0 - 1$. Using $\mathbf{z}_n^\star$ defined by (4.4) we have that

$$\begin{aligned}\sup_{(\pi,P)\in\Theta^{k_0-1}} \mathbb{P}_{\pi,P}(\mathbf{x}_{n\times n}) \;&\leq\; \sum_{\mathbf{z}_n\in[k_0-1]^n} \sup_{(\pi,P)\in\Theta^{k_0-1}} \mathbb{P}_{\pi,P}(\mathbf{x}_{n\times n}, \mathbf{z}_n)\\ &\leq\; (k_0-1)^n \sup_{(\pi,P)\in\Theta^{k_0-1}} \mathbb{P}_{\pi,P}(\mathbf{z}_n^\star, \mathbf{x}_{n\times n})\end{aligned}$$

and on the other hand

$$\begin{aligned}\sup_{(\pi,P)\in\Theta^{k_0}} \mathbb{P}_{\pi,P}(\mathbf{x}_{n\times n}) \;&=\; \sup_{(\pi,P)\in\Theta^{k_0}} \sum_{\bar{\mathbf{z}}_n\in[k_0]^n} \mathbb{P}_{\pi,P}(\bar{\mathbf{z}}_n, \mathbf{x}_{n\times n})\\ &\geq\; \sup_{(\pi,P)\in\Theta^{k_0}} \mathbb{P}_{\pi,P}(\mathbf{z}_n, \mathbf{x}_{n\times n})\,.\end{aligned}$$

Therefore

$$\begin{aligned}\log \frac{\sup_{(\pi,P)\in\Theta^{k_0}} \mathbb{P}_{\pi,P}(\mathbf{x}_{n\times n})}{\sup_{(\pi,P)\in\Theta^{k_0-1}} \mathbb{P}_{\pi,P}(\mathbf{x}_{n\times n})} \;&\geq\; \log \frac{\sup_{(\pi,P)\in\Theta^{k_0}} \mathbb{P}_{\pi,P}(\mathbf{z}_n, \mathbf{x}_{n\times n})}{\sup_{(\pi,P)\in\Theta^{k_0-1}} \mathbb{P}_{\pi,P}(\mathbf{z}_n^\star, \mathbf{x}_{n\times n})}\\ &\quad - n\log(k_0-1)\,.\end{aligned} \tag{4.24}$$

Using that $n\rho_n \to \infty$ on both regimes $\rho_n = \rho > 0$ (dense regime) and $\rho_n \to 0$ (sparse regime) we have, by Lemmas 4.6 and 4.7 that almost surely (4.23) holds for $k = k_0 - 1$. To complete the proof, let $k < k_0 - 1$. In this case we can write

$$\begin{aligned}\log \frac{\sup_{(\pi,P)\in\Theta^{k_0}} \mathbb{P}_{\pi,P}(\mathbf{x}_{n\times n})}{\sup_{(\pi,P)\in\Theta^{k}} \mathbb{P}_{\pi,P}(\mathbf{x}_{n\times n})} \;&=\; \log \frac{\sup_{(\pi,P)\in\Theta^{k_0}} \mathbb{P}_{\pi,P}(\mathbf{x}_{n\times n})}{\sup_{(\pi,P)\in\Theta^{k_0-1}} \mathbb{P}_{\pi,P}(\mathbf{x}_{n\times n})}\\ &\quad + \log \frac{\sup_{(\pi,P)\in\Theta^{k_0-1}} \mathbb{P}_{\pi,P}(\mathbf{x}_{n\times n})}{\sup_{(\pi,P)\in\Theta^{k}} \mathbb{P}_{\pi,P}(\mathbf{x}_{n\times n})}\,.\end{aligned}$$

The first term in the right-hand side can be handled in the same way as in (4.24). On the other hand the second term is non-negative because the maximum likelihood function is a non-decreasing function of the dimension of the model and $k < k_0 - 1$. This finishes the proof of Proposition 4.5 □

5 Discussion

In this paper we introduced a model selection procedure based on the Krichevsky-Trofimov mixture distribution for the number of communities in the Stochastic Block Model. We proved the almost sure convergence (strong consistency) of the penalized estimator (3.1) to the underlying number of communities, without assuming a known upper bound on that quantity. To our knowledge this is the first strong consistency result for an estimator of the number of communities, even in the bounded case.

The family of penalty functions of the form (3.5) are of smaller order compared to the ones used in Wang *et al.* (2017), therefore our results also apply to their family of penalty functions. Moreover, we consider a wider family of sparse models with edge probability of order ρ_n, where ρ_n can decrease to 0 at a rate $n\rho_n \to \infty$. It remains open if it is even possible to obtain consistency in the sparse regime with $\rho_n = 1/n$ and which are the smallest penalty functions for a consistent estimator of the number of communities.

A Proofs of auxiliary results

We begin by stating without proof a basic inequality for the Gamma function. The proof of this result can be found in Davisson *et al.* (1981).

Lemma A.1. *For integers $n = n_1 + \cdots + n_J$ we have that*

$$\frac{\prod_{j=1}^{J} \left(\frac{n_j}{n}\right)^{n_j}}{\prod_{j=1}^{J} \Gamma\left(n_j + \frac{1}{2}\right)} \le \frac{1}{\Gamma\left(n + \frac{1}{2}\right) \Gamma\left(\frac{1}{2}\right)^{J-1}} . \tag{A.1}$$

We now follow by presenting the proof a Proposition 3.1.

Proof of Proposition 3.1. The proof is based on Liu & Narayan (1994, Lemma 3.4). For $(\pi, P) \in \Theta^k$ we have that

$$\mathbb{P}_{\pi,P}(\mathbf{z}_n) = \prod_{a=1}^{k} \pi_a^{n_a} \tag{A.2}$$

and

$$\mathbb{P}_{\pi,P}(\mathbf{x}_{n\times n}|\mathbf{z}_n) = \prod_{1\le a\le b\le k} \left[P_{a,b}^{\tilde{O}_{a,b}} (1 - P_{a,b})^{(\tilde{n}_{a,b} - \tilde{O}_{a,b})} \right]^{\frac{1}{2}} , \tag{A.3}$$

where

$$\tilde{n}_{a,b} = \begin{cases} 2n_{a,b} \,, & 1 \le a,b \le k \,;\, a \ne b \\ n_{a,b} & 1 \le a,b \le k \,;\, a = b \end{cases}$$

and

$$\tilde{O}_{a,b} = \begin{cases} 2O_{a,b} \,, & 1 \le a,b \le k \,;\, a \ne b \\ O_{a,b} & 1 \le a,b \le k \,;\, a = b \,. \end{cases}$$

Using that the maximum likelihood estimators for π_a and $P_{a,b}$ are given by $\dfrac{n_a}{n}$ and $\dfrac{\tilde{O}_{a,b}}{\tilde{n}_{a,b}}$ (respectively) we can bound above (A.2) and (A.3) by

$$\mathbb{P}_{\pi,P}(\mathbf{z}_n) \le \sup_{(\pi,P)\in\Theta^k} \mathbb{P}_{\pi,P}(\mathbf{z}_n) = \prod_{a=1}^{k} \left(\frac{n_a}{n}\right)^{n_a} . \tag{A.4}$$

and

$$\begin{aligned} \mathbb{P}_{\pi,P}(\mathbf{x}_{n\times n}|\mathbf{z}_n) &\le \sup_{(\pi,P)\in\Theta^k} \mathbb{P}_{\pi,P}(\mathbf{x}_{n\times n}|\mathbf{z}_n) \\ &= \prod_{1\le a\le b\le k} \left(\frac{\tilde{O}_{a,b}}{\tilde{n}_{a,b}}\right)^{\tilde{O}_{a,b}/2} \left(1 - \frac{\tilde{O}_{a,b}}{\tilde{n}_{a,b}}\right)^{(\tilde{n}_{a,b} - \tilde{O}_{a,b})/2} . \end{aligned} \tag{A.5}$$

Observe that the Krichevsky-Trofimov mixture distribution defined in (3.3) can be written as

$$\begin{aligned}\mathrm{KT}_k(\mathbf{x}_{n\times n}) &= \sum_{\mathbf{z}_n\in[k]^n}\left(\int_{\Theta_1^k}\mathbb{P}_{\pi,P}(\mathbf{z}_n)\nu_k^1(\pi)d\pi\right)\left(\int_{\Theta_2^k}\mathbb{P}_{\pi,P}(\mathbf{x}_{n\times n}|\mathbf{z}_n)\nu_k^2(P)dP\right)\\ &= \sum_{\mathbf{z}_n\in[k]^n}\mathrm{KT}_k(\mathbf{z}_n)\mathrm{KT}_k(\mathbf{x}_{n\times n}|\mathbf{z}_n)\,,\end{aligned}\tag{A.6}$$

where

$$\nu_k^1(\pi) = \frac{\Gamma(\frac{k}{2})}{\Gamma(\frac{1}{2})^k}\prod_{a=1}^{k}\pi_a^{-1/2}\,,\quad \nu_k^2(P) = \prod_{1\leq a\leq b\leq k}\frac{1}{\Gamma(\frac{1}{2})^2}P_{a,b}^{-1/2}(1-P_{a,b})^{-1/2}\,,$$

$$\Theta_1^k = \{\,\pi\,|\,\pi\in(0,1]^k, \sum_{a=1}^{k}\pi_a = 1\}\,,$$

and

$$\Theta_2^k = \{\,P\,|\,P\in[0,1]^{k\times k}, P \text{ is symmetric}\,\}.$$

We start with the evaluation of $\mathrm{KT}_k(\mathbf{z}_n)$. By a simple calculation we have that

$$\mathrm{KT}_k(\mathbf{z}_n) \;=\; \frac{\Gamma\left(\frac{k}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^k}\,\frac{\prod_{a=1}^k\Gamma\left(n_a+\frac{1}{2}\right)}{\Gamma\left(n+\frac{k}{2}\right)}\,.\tag{A.7}$$

Then combining (A.4) and (A.7) we obtain the bound

$$\frac{\mathbb{P}_{\pi,P}(\mathbf{z}_n)}{\mathrm{KT}_k(\mathbf{z}_n)} \;\leq\; \frac{\Gamma\left(\frac{1}{2}\right)^k\Gamma\left(n+\frac{k}{2}\right)}{\Gamma\left(\frac{k}{2}\right)}\prod_{a=1}^{k}\frac{\left(\frac{n_a}{n}\right)^{n_a}}{\Gamma\left(n_a+\frac{1}{2}\right)}\tag{A.8}$$

Using the fact that $n_1+\cdots+n_k = n$ and Lemma A.1 we can bound the second factor in the right-hand side of the last inequality by

$$\prod_{a=1}^{k}\frac{\left(\frac{n_a}{n}\right)^{n_a}}{\Gamma\left(n_a+\frac{1}{2}\right)} \;\leq\; \frac{1}{\Gamma\left(\frac{1}{2}\right)^{k-1}\,\Gamma\left(n+\frac{1}{2}\right)}\tag{A.9}$$

and then we obtain the bound

$$\frac{\mathbb{P}_{\pi,P}(\mathbf{z}_n)}{\mathrm{KT}_k(\mathbf{z}_n)} \;\leq\; \frac{\Gamma\left(\frac{1}{2}\right)\Gamma\left(n+\frac{k}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(n+\frac{1}{2}\right)}\,.\tag{A.10}$$

The same arguments above can be used to derive the upper bound

$$\frac{\mathbb{P}_{\pi,P}(\mathbf{x}_{n\times n}|z_1^n)}{\mathrm{KT}_k(\mathbf{x}_{n\times n}|z_1^n)} \;\leq\; \Gamma\left(\tfrac{1}{2}\right)^{\frac{k(k-1)}{2}+k}\prod_{1\leq a\leq b\leq k}\frac{\Gamma\left(\frac{\tilde{n}_{a,b}}{2}+1\right)}{\Gamma\left(\frac{\tilde{n}_{a,b}}{2}+\frac{1}{2}\right)}\,.\tag{A.11}$$

Thus combining the bounds in (A.10) and (A.11) we obtain

$$\log \frac{\mathbb{P}_{\pi,P}(\mathbf{x}_{n\times n})}{\mathrm{KT}_k(\mathbf{x}_{n\times n})} \;\leq\; \log\left(\frac{\Gamma(\frac{1}{2})\Gamma(n+\frac{k}{2})}{\Gamma(\frac{k}{2})\Gamma(n+\frac{1}{2})}\right) + \left(\frac{k(k-1)}{2}+k\right)\log\Gamma\left(\tfrac{1}{2}\right) + \sum_{1\leq a\leq b\leq k}\log\left(\frac{\Gamma\left(\frac{\tilde{n}_{a,b}}{2}+1\right)}{\Gamma\left(\frac{\tilde{n}_{a,b}}{2}+\frac{1}{2}\right)}\right). \tag{A.12}$$

It can be shown, as in Gassiat & Boucheron (2003, Appendix I), that

$$\log\left(\frac{\Gamma(\frac{1}{2})\Gamma(n+\frac{k}{2})}{\Gamma(\frac{k}{2})\Gamma(n+\frac{1}{2})}\right) \;\leq\; \left(\tfrac{k-1}{2}\right)\log n + \frac{k(k-1)}{4n} + \frac{1}{12n} + \log\frac{\Gamma(\frac{1}{2})}{\Gamma(\frac{k}{2})} \tag{A.13}$$

and for $1 \leq a \leq b \leq k$,

$$\begin{aligned}\log\left(\frac{\Gamma\left(\frac{\tilde{n}_{a,b}}{2}+1\right)}{\Gamma\left(\frac{\tilde{n}_{a,b}}{2}+\frac{1}{2}\right)}\right) \;&\leq\; \tfrac{1}{2}\log\left(\tfrac{\tilde{n}_{a,b}}{2}\right) + \frac{1}{n_{a,b}} + \frac{1}{6n_{a,b}} \\ &\leq\; \log n + \tfrac{7}{6}\end{aligned} \tag{A.14}$$

where the last inequality follows from the fact that $\tilde{n}_{a,b} \leq 2n^2$ for all a and b. Setting

$$c_{k,n} = \frac{k(k+1)}{2}\log\Gamma\left(\tfrac{1}{2}\right) + \frac{k(k-1)}{4n} + \frac{1}{12n} + \log\frac{\Gamma(\frac{1}{2})}{\Gamma(\frac{k}{2})} + \frac{7k(k+1)}{12}$$

and combining (A.12), (A.13) and (A.14) the result follows. $\square$

Now we state and prove a lemma that is useful to bound the probability of overestimation.

Lemma A.2. *For $k > k_0$ we have*

$$\mathbb{P}_{\pi^0,P^0}(\hat{k}_{KT}(\mathbf{x}_{n\times n}) = k) \;\leq\; \exp\left\{\frac{(k_0(k_0+2)-1)}{2}\log n + c_{k_0,n} + d_{k_0,k,n}\right\},$$

where $d_{k_0,k,n} = pen(k_0,n) - pen(k,n)$.

Proof. For $k > k_0$ define the events

$$A(k) \;=\; \left\{\arg\max_{k'}\{\,\log\mathrm{KT}_{k'}(\mathbf{x}_{n\times n}) - \mathrm{pen}(k',n)\,\} = k\right\}$$

and

$$B(k) \;=\; \left\{\mathrm{KT}_{k_0}(\mathbf{x}_{n\times n}) \leq \mathrm{KT}_k(\mathbf{x}_{n\times n})\mathrm{e}^{d_{k_0,k,n}}\right\}.$$

Observe that $A(k) \subset B(k)$ for all k. Then

$$\begin{aligned}\mathbb{P}_{\pi^0,P^0}(\hat{k}_{\mathrm{KT}}(\mathbf{x}_{n\times n}) = k) \;&=\; \sum_{\mathbf{x}_{n\times n}} \mathbb{P}_{\pi^0,P^0}(\mathbf{x}_{n\times n})\mathbb{1}_{A(k)} \\ &\leq\; \sum_{\mathbf{x}_{n\times n}} \mathbb{P}_{\pi^0,P^0}(\mathbf{x}_{n\times n})\mathbb{1}_{B(k)}\,.\end{aligned} \tag{A.15}$$

By Proposition 3.1

$$\log \mathbb{P}_{\pi^0,P^0}(\mathbf{x}_{n\times n}) \;\leq\; \log \sup_{(\pi,P)\in\Theta^{k_0}} \mathbb{P}_{\pi,P}(\mathbf{x}_{n\times n})$$
$$\leq\; \log \mathrm{KT}_{k_0}(\mathbf{x}_{n\times n}) + \left(\tfrac{k_0(k_0+2)-1}{2}\right)\log n + c_{k_0,n}$$

and therefore

$$\mathbb{P}_{\pi^0,P^0}(\mathbf{x}_{n\times n}) \;\leq\; \mathrm{KT}_{k_0}(\mathbf{x}_{n\times n}) n^{\left(\frac{k_0(k_0+2)-1}{2}\right)} e^{c_{k_0,n}}\,. \tag{A.16}$$

Applying (A.16) in (A.15) we obtain that

$$\mathbb{P}_{\pi^0,P^0}(\hat{k}_{\mathrm{KT}}(\mathbf{x}_{n\times n})=k) \;\leq\; \sum_{\mathbf{x}_{n\times n}} \mathrm{KT}_{k_0}(\mathbf{x}_{n\times n}) n^{\left(\frac{k_0(k_0+2)-1}{2}\right)} e^{c_{k_0,n}} \mathbb{1}_{B(k)}$$
$$\leq\; \sum_{\mathbf{x}_{n\times n}} \mathrm{KT}_{k}(\mathbf{x}_{n\times n})\, e^{d_{k_0,k,n}} n^{\left(\frac{k_0(k_0+2)-1}{2}\right)} e^{c_{k_0,n}}$$
$$= \exp\left\{ \tfrac{(k_0(k_0+2)-1)}{2}\log n + c_{k_0,n} + d_{k_0,k,n} \right\}\,,$$

where the last equality follows from the fact that $\mathrm{KT}_k(\cdot)$ is a probability distribution on the space of adjacency matrices. This concludes the proof of Lemma A.2. □

References

Amini, Arash A, Chen, Aiyou, Bickel, Peter J, Levina, Elizaveta, *et al.* . 2013. Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, **41**(4), 2097–2122.

Bickel, Peter, Choi, David, Chang, Xiangyu, & Zhang, Hai. 2013. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *The Annals of Statistics*, 1922–1943.

Bickel, Peter J, & Chen, Aiyou. 2009. A nonparametric view of network models and Newman–Girvan and other modularities. *Proceedings of the National Academy of Sciences*, **106**(50), 21068–21073.

Csiszar, I., & Talata, Z. 2006. Context tree estimation for not necessarily finite memory processes, via BIC and MDL. *IEEE Transactions on Information Theory*, **52**(3), 1007–1016.

Daudin, J-J, Picard, Franck, & Robin, Stéphane. 2008. A mixture model for random graphs. *Statistics and computing*, **18**(2), 173–183.

Davisson, L, McEliece, R, Pursley, M, & Wallace, Mark. 1981. Efficient universal noiseless source codes. *IEEE Transactions on Information Theory*, **27**(3), 269–279.

Gassiat, Elisabeth, & Boucheron, Stéphane. 2003. Optimal error exponents in hidden Markov models order estimation. *IEEE Transactions on Information Theory*, **49**(4), 964–980.

Holland, Paul W, Laskey, Kathryn Blackmond, & Leinhardt, Samuel. 1983. Stochastic blockmodels: First steps. *Social networks*, **5**(2), 109–137.

Hu, Jianwei, Qin, Hong, Yan, Ting, & Zhao, Yunpeng. 2016. On consistency of model selection for stochastic block models. *arXiv preprint arXiv:1611.01238*.

Krichevsky, Raphail E., & Trofimov, Victor K. 1981. The performance of universal encoding. *IEEE Trans. Information Theory*, **27**(2), 199–206.

Latouche, Pierre, Birmele, Etienne, & Ambroise, Christophe. 2012. Variational Bayesian inference and complexity control for stochastic block models. *Statistical Modelling*, **12**(1), 93–115.

Le, Can M, & Levina, Elizaveta. 2015. Estimating the number of communities in networks by spectral methods. *arXiv preprint arXiv:1507.00827*.

Liu, Chuang-Chun, & Narayan, Prakash. 1994. Order estimation and sequential universal data compression of a hidden Markov source by the method of mixtures. *IEEE Transactions on Information Theory*, **40**(4), 1167–1180.

Rohe, Karl, Chatterjee, Sourav, Yu, Bin, *et al.* . 2011. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, **39**(4), 1878–1915.

Su, Liangjun, Wang, Wuyi, & Zhang, Yichong. 2017. Strong Consistency of Spectral Clustering for Stochastic Block Models. *arXiv preprint arXiv:1710.06191*.

van der Pas, SL, van der Vaart, AW, *et al.* . 2017. Bayesian Community Detection. *Bayesian Analysis*.

Wang, YX Rachel, Bickel, Peter J, *et al.* . 2017. Likelihood-based model selection for stochastic block models. *The Annals of Statistics*, **45**(2), 500–528.