Electronic Journal of Statistics
Vol. 14 (2020) 1234–1268
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1695

Consistency and asymptotic normality of Latent Block Model estimators

Vincent Brault

Univ. Grenoble Alpes, CNRS, Grenoble INP[*]*, LJK, 38000 Grenoble, France*
e-mail: vincent.brault@univ-grenoble-alpes.fr
url: https://www-ljk.imag.fr/membres/Vincent.Brault/

and

Christine Keribin

Université Paris-Saclay, CNRS, Inria, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France
e-mail: christine.keribin@math.u-psud.fr
url: https://www.math.u-psud.fr/~keribin/

and

Mahendra Mariadassou

MaIAGE, INRAE, Université Paris-Saclay, 78352 Jouy-en-Josas, France
e-mail: mahendra.mariadassou@inrae.fr
url: https://mahendra-mariadassou.github.io/

Abstract: The Latent Block Model (LBM) is a model-based method to cluster simultaneously the d columns and n rows of a data matrix. Parameter estimation in LBM is a difficult and multifaceted problem. Although various estimation strategies have been proposed and are now well understood empirically, theoretical guarantees about their asymptotic behavior is rather sparse and most results are limited to the binary setting. We prove here theoretical guarantees in the valued settings. We show that under some mild conditions on the parameter space, and in an asymptotic regime where $\log(d)/n$ and $\log(n)/d$ tend to 0 when n and d tend to infinity, (1) the maximum-likelihood estimate of the complete model (with known labels) is consistent and (2) the log-likelihood ratios are equivalent under the complete and observed (with unknown labels) models. This equivalence allows us to transfer the asymptotic consistency, and under mild conditions, asymptotic normality, to the maximum likelihood estimate under the observed model. Moreover, the variational estimator is also consistent and, under the same conditions, asymptotically normal.

Keywords and phrases: Latent Block Model, asymptotic normality, Maximum Likelihood Estimate, Concentration Inequality.

Received July 2019.

[*]Institute of Engineering Univ. Grenoble Alpes

1. Introduction

Co-clustering is an unsupervised method to cluster simultaneously the n rows and d columns of a rectangular data matrix. The assignments of each row to one of the row-clusters and of each column to one of the column-clusters are unknown and the aim is to determine them. Then, rows and columns can be re-ordered according to their assignments, highlighting the natural structure of the data with distinct blocks having homogeneous observations. This leads to a parsimonious data representation, as can be shown on Figure 1.

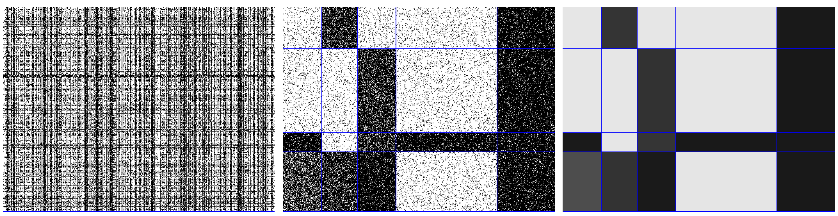

FIG 1. *A binary data matrix before (left) and after (middle) row and column reordering, and its parsimonious data representation (right).*

Co-clustering can be used in numerous applications, and especially ones with large data sets, such as recommendation systems (to discover a segmentation of customers with regard to a segmentation of products), genomics (to simultaneously define groups of genes having the same expression with regards to groups of experimental conditions) or text mining (to define simultaneously groups of texts and groups of words).

Among the co-clustering methods, the Latent Block Model (LBM) defines a probabilistic model as a mixture model with latent rows and columns assignments. LBM can deal with binary [6], Gaussian [8], categorical [9] or count [7] data. Due to the complex dependency structure induced by this modeling, neither the likelihood, nor the distribution of the assignments conditionally to the observations needed in the E-step of the EM algorithm, traditionnally used for mixture models, are numerically tractable. Estimation can be however performed either with a variational approximation leading to an approximate value of the maximum likelihood estimator, or with a Bayesian approach (VBayes algorithm or Gibbs sampler). For example, [9] recommends using a Gibbs sampler combined with a VBayes algorithm.

The asymptotics of the maximum likelihood (MLE) and variational (VE) estimators also raise interesting theoretical questions. This topic was first addressed for the stochastic block model (SBM) [13], where the data is a random graph encoded by its adjacency binary matrix: the rows and columns represent the nodes, so that there is only one partition, shared by rows and columns, and a unique asymptotic direction.

For a binary SBM and under the true parameter value, Theorem 3 of [4] states that the distribution of the assignments conditionally to the observations converges to a Dirac of the real assignments. Moreover, this convergence remains valid under the estimated parameter value, assuming that this estimator converges at rate at least n^{-1}, where n is the number of nodes (Proposition 3.8). This assumption is not trivial, and it was not established that such an estimator exists except in some particular cases ([1] for example). [10] presented a unified frame for LBM and SBM in case of valued observations satisfying a concentration inequality, and showed the consistency of the conditional distribution of the assignments under all parameter values in a neighborhood of the true value. [3] and [2] proved the consistency and asymptotic normality of the MLE for the binary SBM but failed to account for complications induced by symmetries in the parameter. Building upon the work from [4], they first studied the asymptotic behavior of the MLE in the complete model (observations and assignments) with binary observations which is simple to handle; then, they showed that the complete likelihood and the marginal likelihood have similar asymptotic behaviors by the use of a Bernstein inequality for bounded observations.

Following the main ideas of [2], we prove that the observed likelihood ratio and the complete likelihood ratio computed at the true assignments are asymptotically equivalent, up to a multiplicative term. This term depends on some model symmetry and was omitted in [2] although it is necessary to prove the asymptotic results. We then settle the asymptotic normality of the maximum likelihood and variational estimators. All these results are stated not only for binary observations, but also more generally for observations coming from univariate exponential families in canonical form, which is essential regarding the LBM usages. This leads us to develop a Bernstein-type inequality for sub-exponential variables as the Hoeffing's concentration inequality used in [2] is only relevant for upper-bounded observations.

The paper is organized as follows. The model, main assumptions and notations are introduced in Section 2, where the concept of model symmetry is also discussed. Section 3 proves the asymptotic normality of the complete likelihood estimator, and section 4 studies conditional and profile log-likelihoods. Our main result showing that the observed likelihood ratio behaves like the complete likelihood ratio is stated in section 5, and its consequences in terms of consistency and asymptotic normality of the MLE and variational estimators are presented in section 6. Most of the proofs are postponed to the appendices to improve the general readibility: appendix A for properties of conditional and profile log-likelihoods, B for the steps of the main result, C for concentration inequalities for specific sub-exponential variables and D for other technical results.

2. Model, assumptions and definitions

We observe a data matrix $X = (x_{ij})$ with n rows and d columns. The LBM assumes that there exists a latent structure in the form of the Cartesian product of a partition of g row-clusters by a partition of m column-clusters with the following characteristics:

- the latent row assignments $\mathbf{z} = (\mathbf{z}_1, \dots, \mathbf{z}_n)$ are independent and identically distributed with a common multinomial distribution on g categories:

$$\mathbf{z}_i \sim \mathcal{M}(1, \boldsymbol{\pi} = (\pi_1, \dots, \pi_g))$$

  For $k = 1, \dots, g$, $z_{ik} = 1$ if row i belongs to row-group k, 0 otherwise. In the same way, the latent column assignments $\mathbf{w} = (\mathbf{w}_1, \dots, \mathbf{w}_d)$ are i.i.d. multinomial variables with m categories:

$$\mathbf{w}_j \sim \mathcal{M}(1, \boldsymbol{\rho} = (\rho_1, \dots, \rho_m))$$

  For $\ell = 1, \dots, m$, $w_{j\ell} = 1$ if column j belongs to column-group ℓ and 0 otherwise.
- the row and column assignments are independent: $p(\mathbf{z}, \mathbf{w}) = p(\mathbf{z})p(\mathbf{w})$
- conditionally to row and column assignments $\mathbf{z} \times \mathbf{w}$, the observed data X_{ij} are independent, and their conditional distribution $\varphi(., \alpha)$ belongs to the same parametric family, which parameter α only depends on the given block:

$$X_{ij} | \{z_{ik} w_{j\ell} = 1\} \sim \varphi(., \alpha_{k\ell}).$$

Hence, the complete parameter set is $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha}) \in \boldsymbol{\Theta}$, with $\boldsymbol{\alpha} = (\alpha_{11}, \dots, \alpha_{gm})$ and $\boldsymbol{\Theta}$ the parameter space. Figure 2 summarizes these notations.

Remark 2.1. Group, class and cluster in one hand, label and assignment in the other hand will be used indistinctly. Moreover, for notation convenience, $\sum_i$, $\sum_j$, $\sum_k$, $\sum_\ell$ stand for $\sum_{i=1}^n$, $\sum_{j=1}^d$, $\sum_{k=1}^g$, $\sum_{\ell=1}^m$.

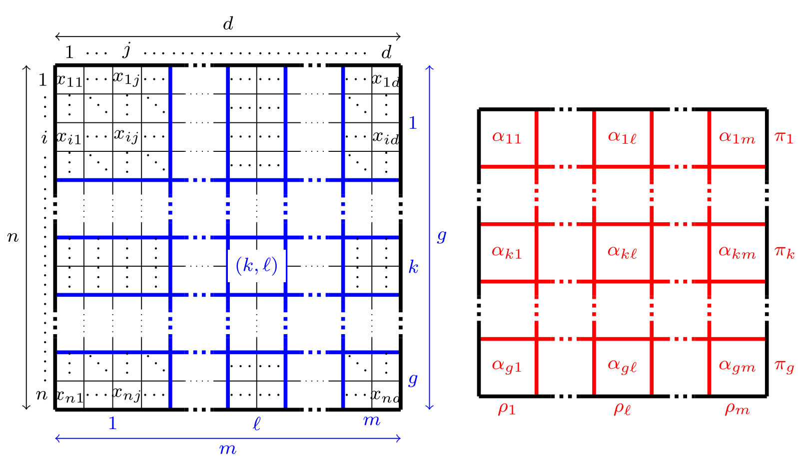

FIG 2. *Notations. Left: Notations for the elements of observed data matrix are in black, notations for the block clusters are in blue. Right: Notations for the model parameter.*

When performing inference from data, we denote $\boldsymbol{\theta}^\star = (\boldsymbol{\pi}^\star, \boldsymbol{\rho}^\star, \boldsymbol{\alpha}^\star)$ the true parameter set, *i.e.* the parameter values used to generate the data, and $\mathbf{z}^\star$ and $\mathbf{w}^\star$ the true (and usually unobserved) row and column assignments. For indicator membership variables $\mathbf{z}$ and $\mathbf{w}$, we also denote:

- $z_{+k} = \sum_i z_{ik}$ and $w_{+\ell} = \sum_j w_{j\ell}$
- $z^\star_{+k}$ and $w^\star_{+\ell}$ their counterpart for $\mathbf{z}^\star$ and $\mathbf{w}^\star$.

The confusion matrix allows one to compare the partitions.

Definition 2.2 (confusion matrices)**.** *For given assignments* $\mathbf{z}$ *and* $\mathbf{z}^\star$ *(resp.* $\mathbf{w}$ *and* $\mathbf{w}^\star$*), we define the* confusion matrix *between* $\mathbf{z}$ *and* $\mathbf{z}^\star$ *(resp.* $\mathbf{w}$ *and* $\mathbf{w}^\star$*), denoted* $\mathbb{R}_g(\mathbf{z})$ *(resp.* $\mathbb{R}_m(\mathbf{w})$*), as follows:*

$$\mathbb{R}_g(\mathbf{z})_{kk'} = \frac{1}{n}\sum_i z^\star_{ik} z_{ik'} \quad \textit{and} \quad \mathbb{R}_m(\mathbf{w})_{\ell\ell'} = \frac{1}{d}\sum_j w^\star_{j\ell} w_{j\ell'}$$

2.1. Likelihood

When the labels are known, the *complete log-likelihood* is given by:

$$\begin{aligned} \mathcal{L}_c(\mathbf{z}, \mathbf{w}; \boldsymbol{\theta}) &= \log p(\mathbf{x}, \mathbf{z}, \mathbf{w}; \boldsymbol{\theta}) \\ &= \log\left\{\left(\prod_{i,k} \pi_k^{z_{ik}}\right)\left(\prod_{j,\ell} \rho_\ell^{w_{j\ell}}\right)\left(\prod_{i,j,k,\ell} \varphi\left(x_{ij}; \alpha_{k\ell}\right)^{z_{ik}w_{j\ell}}\right)\right\} \\ &= \log\left\{\left(\prod_i \pi_{z_i}\right)\left(\prod_j \rho_{w_j}\right)\left(\prod_{i,j} \varphi\left(x_{ij}; \alpha_{z_i w_j}\right)\right)\right\}. \end{aligned} \tag{2.1}$$

In an unsupervised setting, the labels are unobserved and the *observed log-likelihood* is obtained by marginalization over all the label configurations:

$$\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{x}; \boldsymbol{\theta}) = \log\left(\sum_{\mathbf{z}\in\mathcal{Z}, \mathbf{w}\in\mathcal{W}} p(\mathbf{x}, \mathbf{z}, \mathbf{w}; \boldsymbol{\theta})\right).$$

Due to the double missing data structure $\mathbf{z}$ for rows and $\mathbf{w}$ for columns, neither the observed likelihood nor the E-step of the EM algorithm are tractable. Estimation can can nevertheless be performed either by numerical approximation, or by MCMC methods [see 8, 9].

2.2. Assumptions

We focus here on LBM where φ belongs to a regular univariate exponential family set in canonical form:

$$\varphi(x, \alpha) = b(x)\exp(\alpha x - \psi(\alpha)),$$

The canonical parameter α belongs to a space $\mathcal{A}$, so that $\varphi(\cdot, \alpha)$ is well defined for all $\alpha \in \mathcal{A}$. Classical properties of exponential families ensure that ψ is convex, infinitely differentiable on $\mathring{\mathcal{A}}$, and $(\psi')^{-1}$ is well defined on $\psi'(\mathring{\mathcal{A}})$.

When $X_\alpha \sim \varphi(.,\alpha)$,

$$\mathbb{E}[X_\alpha] = \psi'(\alpha) \text{ and } \mathbb{V}[X_\alpha] = \psi''(\alpha).$$

Notice that the definition of the exponential family used here relies on an exhaustive statistic that is X itself. This for a simple convenience. Family sets of the form $\varphi(x,\alpha) = b(x)\exp(\alpha t(x) - \psi(\alpha))$ can also be considered, all the further developments as Bernstein and concentration inequalities then concerning the exhaustive statistics $t(X)$.

Moreover, we make the following assumptions on the parameter space:

H_1: There exists a positive constant c, and a compact C_α such that

$$\boldsymbol{\Theta} \subset [c, 1-c]^g \times [c, 1-c]^m \times C_\alpha^{g\times m} \quad \text{with} \quad C_\alpha \subset \mathring{\mathcal{A}}.$$

H_2: The true parameter $\boldsymbol{\theta}^\star = (\boldsymbol{\pi}^\star, \boldsymbol{\rho}^\star, \boldsymbol{\alpha}^\star)$ lies in the relative interior of $\boldsymbol{\Theta}$.

H_3: The mixture measure of LBM is identifiable: $\boldsymbol{\theta}^\star$ is identifiable up to a permutation of the row-labels and column-labels (see definition 2.7 of equivalent parameters).

The previous assumptions are standard. Notice that the following conditions are necessary for H_3 to hold:

H_{3a}: The map $\alpha \mapsto \varphi(\cdot,\alpha)$ is injective.

H_{3b}: Each row and each column of $\boldsymbol{\alpha}^\star$ is unique.

[9] gives sufficient conditions for the generic identifiability of the categorical LBM, *i.e.* except on a manifold set of null Lebesgue measure in $\boldsymbol{\Theta}$ and this property is easily extended to the case of observations from a univariate exponential family. For binary SBM, [2] added the assumption $p \geq (\log n)/n$ on the parameter p of the Bernoulli distribution to take into account sparsity.

Assumption H_1 ensures that the group proportions π_k and ρ_ℓ are bounded away from 0 and 1 so that no group disappears when n and d go to infinity. It also ensures that α is bounded away from the boundaries of $\mathcal{A}$ and that there exists a positive value $\kappa > 0$, such that $[\alpha-\kappa, \alpha+\kappa] \subset \mathring{\mathcal{A}}$ for all parameters α of C_α, which is essential to prove a uniform Bernstein inequality on the X.

Moreover, we define the quantity $\delta(\boldsymbol{\alpha})$ that captures the separation between row-groups or column-groups: low values of $\delta(\boldsymbol{\alpha})$ mean that two row-classes or two column-classes are very similar.

Definition 2.3 (class distinctness)**.** *For $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha}) \in \boldsymbol{\Theta}$. We define:*

$$\delta(\boldsymbol{\alpha}) = \min\left\{ \min_{\ell\neq\ell'} \max_k \mathrm{KL}(\alpha_{k\ell}, \alpha_{k\ell'}), \min_{k\neq k'} \max_\ell \mathrm{KL}(\alpha_{k\ell}, \alpha_{k'\ell}) \right\}$$

with $\mathrm{KL}(\alpha,\alpha') = \mathbb{E}_\alpha[\log(\varphi(X,\alpha)/\varphi(X,\alpha'))] = \psi'(\alpha)(\alpha-\alpha') + \psi(\alpha') - \psi(\alpha)$ *the Kullback divergence between* $\varphi(.,\alpha)$ *and* $\varphi(.,\alpha')$.

Remark 2.4. Since $\boldsymbol{\alpha}^\star$ has distinct rows and distinct columns (H_3), $\delta(\boldsymbol{\alpha}^\star) > 0$.

Remark 2.5. These assumptions are satisfied for many distributions, including but not limited to:

- Bernoulli, when the proportion p is bounded away from 0 and 1, or natural parameter $\alpha = \log(p/(1-p))$ bounded away from $\pm\infty$;
- Poisson, when the mean λ is bounded away from 0 and $+\infty$, or natural parameter $\alpha = \log(\lambda)$ bounded away from $\pm\infty$;
- Gaussian with known variance when the mean μ, which is also the natural parameter, is bounded away from $\pm\infty$.

In particular, the conditions stating that ψ is twice differentiable and that $(\psi')^{-1}$ exists are equivalent to assuming that X_α has positive and finite variance for all values of α in the parameter space.

2.3. Model symmetry

The LBM is a generalized mixture model and as such is subject to label switching. Moreover, the study of the asymptotics will involve the complete likelihood where symmetry properties on the parameter must be taken into account. We first recall the definition of a permutation in LBM, then define equivalence relationships for assignments and parameter, and discuss model symmetry.

Definition 2.6 (permutation)**.** *Let s be a permutation on $\{1,\dots,g\}$ and t a permutation on $\{1,\dots,m\}$. If $\boldsymbol{A}$ is a matrix with g columns, we define $\boldsymbol{A}^s$ as the matrix obtained by permuting the columns of $\boldsymbol{A}$ according to s,* i.e. *for any row i and column k of $\boldsymbol{A}$, $A^s_{ik} = A_{is(k)}$. If $\boldsymbol{B}$ is a matrix with m columns and $\boldsymbol{C}$ is a matrix with g rows and m columns, $\boldsymbol{B}^t$ and $\boldsymbol{C}^{s,t}$ are defined similarly:*

$$\boldsymbol{A}^s = \left(A_{is(k)}\right)_{i,k} \quad \boldsymbol{B}^t = \left(B_{jt(\ell)}\right)_{j,\ell} \quad \boldsymbol{C}^{s,t} = \left(C_{s(k)t(\ell)}\right)_{k,\ell}$$

Definition 2.7 (equivalence)**.** *We define the following equivalence relationships:*

- *Two assignments $(\mathbf{z},\mathbf{w})$ and $(\mathbf{z}',\mathbf{w}')$ are* equivalent*, denoted $\sim$, if they are equal up to label permutation,* i.e. *there exist two permutations s and t such that $\mathbf{z}' = \mathbf{z}^s$ and $\mathbf{w}' = \mathbf{w}^t$.*
- *Two parameters $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ are* equivalent*, denoted $\sim$, if they are equal up to label permutation,* i.e. *there exist two permutations s and t such that $(\boldsymbol{\pi}^s, \boldsymbol{\rho}^t, \boldsymbol{\alpha}^{s,t}) = (\boldsymbol{\pi}', \boldsymbol{\rho}', \boldsymbol{\alpha}')$. This is* label-switching.
- *$(\boldsymbol{\theta},\mathbf{z},\mathbf{w})$ and $(\boldsymbol{\theta}',\mathbf{z}',\mathbf{w}')$ are* equivalent*, denoted $\sim$, if they are equal up to label permutation on $\boldsymbol{\alpha}$,* i.e. *there exist two permutations, s and t such that $(\boldsymbol{\alpha}^{s,t}, \mathbf{z}^s, \mathbf{w}^t) = (\boldsymbol{\alpha}', \mathbf{z}', \mathbf{w}')$.*

The last equivalence relationship is not concerned with $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$. It is useful when dealing with the conditional likelihood $p(\mathbf{x}|\mathbf{z},\mathbf{w};\boldsymbol{\theta})$ which depends neither on $\boldsymbol{\pi}$ nor $\boldsymbol{\rho}$: in fact, if $(\boldsymbol{\theta},\mathbf{z},\mathbf{w}) \sim (\boldsymbol{\theta}',\mathbf{z}',\mathbf{w}')$, then for all $\mathbf{x}$, we have $p(\mathbf{x}|\mathbf{z},\mathbf{w};\boldsymbol{\theta}) = p(\mathbf{x}|\mathbf{z}',\mathbf{w}';\boldsymbol{\theta}')$. Note also that $\mathbf{z} \sim \mathbf{z}^\star$ (resp. $\mathbf{w} \sim \mathbf{w}^\star$) if and only if there exists a permutation of the rows of the confusion matrix $\mathbb{R}_g(\mathbf{z})$ (resp. $\mathbb{R}_m(\mathbf{w})$) leading to a diagonal matrix.

Definition 2.8 (symmetry)**.** *We say that the parameter* $\boldsymbol{\theta}$ exhibits symmetry for the permutations s, t *if*

$$(\boldsymbol{\pi}^s, \boldsymbol{\rho}^t, \boldsymbol{\alpha}^{s,t}) = (\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha}).$$

$\boldsymbol{\theta}$ exhibits symmetry *if it exhibits symmetry for any non trivial pair of permutations* (s, t)*. Finally the set of pairs* (s, t) *for which* $\boldsymbol{\theta}$ *exhibits symmetry is denoted* $\mathrm{Sym}(\boldsymbol{\theta})$*.*

Remark 2.9. The set of parameters that exhibit symmetry is a manifold of null Lebesgue measure in $\boldsymbol{\Theta}$. This notion of symmetry is subtler than and different from label switching. To emphasize the difference between equivalence and symmetry, consider the following model: $\boldsymbol{\pi} = (1/2, 1/2)$, $\boldsymbol{\rho} = (1/3, 2/3)$ and $\boldsymbol{\alpha} = \left(\begin{smallmatrix} \alpha_1 & \alpha_2 \\ \alpha_2 & \alpha_1 \end{smallmatrix}\right)$ with $\alpha_1 \neq \alpha_2$. The only permutations of interest here are $s = t = [1\ 2]$. Choose any $\mathbf{z}$ and $\mathbf{w}$. Because of label switching, we know that $p(\mathbf{x}, \mathbf{z}^s, \mathbf{w}^t; \boldsymbol{\theta}^{s,t}) = p(\mathbf{x}, \mathbf{z}, \mathbf{w}; \boldsymbol{\theta})$. $(\mathbf{z}^s, \mathbf{w}^t)$ and $(\mathbf{z}, \mathbf{w})$ have the same likelihood but under *different* parameters $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^{s,t}$. If however, $\boldsymbol{\rho} = (1/2, 1/2)$, then $(s, t) \in \mathrm{Sym}(\boldsymbol{\theta})$ and $\boldsymbol{\theta}^{s,t} = \boldsymbol{\theta}$ so that $(\mathbf{z}, \mathbf{w})$ and $(\mathbf{z}^s, \mathbf{w}^t)$ have the same likelihood under the *same* parameter $\boldsymbol{\theta}$. In particular, if $(\mathbf{z}, \mathbf{w})$ is a maximum-likelihood assignment under $\boldsymbol{\theta}$, so is $(\mathbf{z}^s, \mathbf{w}^t)$. In other words, if $\boldsymbol{\theta}$ exhibits symmetry, the maximum-likelihood assignment is *not unique* under the true model and there are at least $\#\,\mathrm{Sym}(\boldsymbol{\theta})$ of them. This has important implications for the asymptotics of the observed likelihood ratio.

2.4. Distance and local assignments

We define the distance up to equivalence between two sets of assignments as follows:

Definition 2.10 (distance)**.** *The distance, up to equivalence, between configurations* $\mathbf{z}$ *and* $\mathbf{z}^\star$ *is defined as*

$$\|\mathbf{z} - \mathbf{z}^\star\|_{0,\sim} = \inf_{\mathbf{z}' \sim \mathbf{z}} \|\mathbf{z}' - \mathbf{z}^\star\|_0$$

where, for all matrix $\mathbf{z}$*,* $\|\cdot\|_0$ *is the Hamming norm*

$$\|\mathbf{z}\|_0 = \sum_{i,k} \mathbb{1}\{z_{ik} \neq 0\}.$$

A similar definition is set for the distance between $\mathbf{w}$ *and* $\mathbf{w}^\star$*.*

This allows us to define a neighborhood of radius r in the assignment space, taking into account equivalent assignments classes.

Definition 2.11 (Set of local assignments)**.** *We denote* $S(\mathbf{z}^\star, \mathbf{w}^\star, r)$ *the set of configurations that have a representative (for* $\sim$*) within relative radius* r *of* $(\mathbf{z}^\star, \mathbf{w}^\star)$*:*

$$S(\mathbf{z}^\star, \mathbf{w}^\star, r) = \{(\mathbf{z}, \mathbf{w}) : \|\mathbf{z} - \mathbf{z}^\star\|_{0,\sim} \leq rn \text{ and } \|\mathbf{w} - \mathbf{w}^\star\|_{0,\sim} \leq rd\}$$

3. Asymptotic properties in the complete data model

As stated in the introduction, we first study the asymptotic properties of the complete data model. Let $\widehat{\boldsymbol{\theta}}_c = (\widehat{\boldsymbol{\pi}}, \widehat{\boldsymbol{\rho}}, \widehat{\boldsymbol{\alpha}})$ be the MLE of $\boldsymbol{\theta}$ in the complete data model, where the real assignments $\mathbf{z} = \mathbf{z}^\star$ and $\mathbf{w} = \mathbf{w}^\star$ are known. We can derive the following general estimates from Equation (2.1):

$$\begin{aligned} \widehat{\pi}_k(\mathbf{z}) = \frac{z_{+k}}{n} \quad & \widehat{\rho}_\ell(\mathbf{w}) = \frac{w_{+\ell}}{d} \\ \widehat{x}_{k\ell}(\mathbf{z},\mathbf{w}) = \frac{\sum_{ij} x_{ij} z_{ik} w_{j\ell}}{z_{+k} w_{+\ell}} \quad & \widehat{\alpha}_{k\ell} = \widehat{\alpha}_{k\ell}(\mathbf{z},\mathbf{w}) = (\psi')^{-1}\left(\widehat{x}_{k\ell}(\mathbf{z},\mathbf{w})\right) \end{aligned} \tag{3.1}$$

Proposition 3.1. *The matrices* $\Sigma_{\boldsymbol{\pi}^\star} = \operatorname{Diag}(\boldsymbol{\pi}^\star) - \boldsymbol{\pi}^\star (\boldsymbol{\pi}^\star)^T$, $\Sigma_{\boldsymbol{\rho}^\star} = \operatorname{Diag}(\boldsymbol{\rho}^\star) - \boldsymbol{\rho}^\star (\boldsymbol{\rho}^\star)^T$ *are semi-definite positive, of rank* $g-1$ *and* $m-1$*, and* $\widehat{\boldsymbol{\pi}}$ *and* $\widehat{\boldsymbol{\rho}}$ *are asymptotically normal:*

$$\sqrt{n}\left(\widehat{\boldsymbol{\pi}}\left(\mathbf{z}^\star\right) - \boldsymbol{\pi}^\star\right) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0, \Sigma_{\boldsymbol{\pi}^\star}) \quad \text{and} \quad \sqrt{d}\left(\widehat{\boldsymbol{\rho}}\left(\mathbf{w}^\star\right) - \boldsymbol{\rho}^\star\right) \xrightarrow[d\to\infty]{\mathcal{D}} \mathcal{N}(0, \Sigma_{\boldsymbol{\rho}^\star}) \tag{3.2}$$

Similarly, let $V(\boldsymbol{\alpha}^\star)$ *be the matrix defined by* $[V(\boldsymbol{\alpha}^\star)]_{k\ell} = 1/\psi''(\alpha^\star_{k\ell})$ *and* $\Sigma_{\boldsymbol{\alpha}^\star} = \operatorname{Diag}^{-1}(\boldsymbol{\pi}^\star) V(\boldsymbol{\alpha}^\star) \operatorname{Diag}^{-1}(\boldsymbol{\rho}^\star)$*. Then:*

$$\sqrt{nd}\left(\widehat{\alpha}_{k\ell}\left(\mathbf{z}^\star, \mathbf{w}^\star\right) - \alpha^\star_{k\ell}\right) \xrightarrow[n,d\to\infty]{\mathcal{D}} \mathcal{N}(0, \Sigma_{\boldsymbol{\alpha}^\star, k\ell}) \quad \text{for all } k, \ell$$

and the components $\widehat{\alpha}_{k,\ell}$ *are independent.*

Proof: Since $\widehat{\boldsymbol{\pi}}\left(\mathbf{z}^\star\right) = \left(\widehat{\pi}_1\left(\mathbf{z}^\star\right), \ldots, \widehat{\pi}_g\left(\mathbf{z}^\star\right)\right)$ (resp. $\widehat{\boldsymbol{\rho}}\left(\mathbf{w}^\star\right)$) is the sample mean of n (resp. d) i.i.d. multinomial random variables with parameters 1 and $\boldsymbol{\pi}^\star$ (resp. $\boldsymbol{\rho}^\star$), a simple application of the central limit theorem (CLT) gives:

$$\Sigma_{\boldsymbol{\pi}^\star, kk'} = \begin{cases} \pi^\star_k(1-\pi^\star_k) & \text{if} \quad k = k' \\ -\pi^\star_k \pi^\star_{k'} & \text{if} \quad k \neq k' \end{cases} \quad \text{and} \quad \Sigma_{\boldsymbol{\rho}^\star, \ell\ell'} = \begin{cases} \rho^\star_\ell(1-\rho^\star_\ell) & \text{if} \quad \ell = \ell' \\ -\rho^\star_\ell \rho^\star_{\ell'} & \text{if} \quad \ell \neq \ell' \end{cases}$$

which proves Equation (3.2) where $\Sigma_{\boldsymbol{\pi}^\star}$ and $\Sigma_{\boldsymbol{\rho}^\star}$ are semi-definite positive of rank $g-1$ and $m-1$.

Similarly, $\psi'\left(\widehat{\alpha}_{k\ell}\left(\mathbf{z}^\star, \mathbf{w}^\star\right)\right)$ is the average of $z^\star_{+k} w^\star_{+\ell} = nd\widehat{\pi}_k\left(\mathbf{z}^\star\right)\widehat{\rho}_\ell\left(\mathbf{w}^\star\right)$ i.i.d. random variables with mean $\psi'\left(\alpha^\star_{k\ell}\right)$ and variance $\psi''\left(\alpha^\star_{k\ell}\right)$. $nd\widehat{\pi}_k\left(\mathbf{z}^\star\right)\widehat{\rho}_\ell\left(\mathbf{w}^\star\right)$ is itself random but $\widehat{\pi}_k\left(\mathbf{z}^\star\right)\widehat{\rho}_\ell\left(\mathbf{w}^\star\right) \xrightarrow[n,d\to+\infty]{} \pi^\star_k \rho^\star_\ell$ almost surely. Therefore, by Slutsky's lemma and the CLT for random sums of random variables [12], we have:

$$\begin{aligned} &\sqrt{nd\pi^\star_k \rho^\star_\ell}\left(\psi'\left(\widehat{\alpha}_{k\ell}\left(\mathbf{z}^\star, \mathbf{w}^\star\right)\right) - \psi'(\alpha^\star_{k\ell})\right) \\ &\qquad = \sqrt{nd\pi^\star_k \rho^\star_\ell}\left(\frac{\sum_{ij} X_{ij} z^\star_{ik} w^\star_{j\ell}}{nd\widehat{\pi}_k\left(\mathbf{z}^\star\right)\widehat{\rho}_\ell\left(\mathbf{w}^\star\right)} - \psi'(\alpha^\star_{k\ell})\right) \\ &\qquad \xrightarrow[n,d\to+\infty]{\mathcal{D}} \mathcal{N}\left(0, \psi''(\alpha^\star_{k\ell})\right) \end{aligned}$$

The differentiability of $(\psi')^{-1}$ and the delta method then gives:

$$\sqrt{nd}\left(\widehat{\alpha}_{k\ell}\left(\mathbf{z}^\star,\mathbf{w}^\star\right)-\alpha_{k\ell}^\star\right)\xrightarrow[n,d\to+\infty]{\mathcal{D}}\mathcal{N}\left(0,\frac{1}{\pi_k^\star\rho_\ell^\star\psi''(\alpha_{k\ell}^\star)}\right)$$

and the independence results from the independence of $\widehat{\alpha}_{k\ell}\left(\mathbf{z}^\star,\mathbf{w}^\star\right)$ and $\widehat{\alpha}_{k'\ell'}\left(\mathbf{z}^\star,\mathbf{w}^\star\right)$ as soon as $k\neq k'$ or $\ell\neq\ell'$, as they involve different sets of independent variables.

□

Moreover, the complete model is locally asymptotically normal (LAN), as stated in the following proposition. Note that the unusual condition for s and t arises from the constraints $\boldsymbol{\pi}^T\mathbf{1}_g=\boldsymbol{\rho}^T\mathbf{1}_m=1$, where $\mathbf{1}_g$ is the vector of size g filled with 1, which must be satisfied even after perturbing $\boldsymbol{\pi}$ (resp. $\boldsymbol{\rho}$) with s (resp. t).

Proposition 3.2 (Local asymptotic normality)**.** *Let $\mathcal{L}_c^\star$ the map defined by $\boldsymbol{\theta}=(\boldsymbol{\pi},\boldsymbol{\rho},\boldsymbol{\alpha})\mapsto\log p\left(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}\right)$ and note $I_{\boldsymbol{\pi}^\star}=\mathrm{Diag}^{-1}(\boldsymbol{\pi}^\star)$, $I_{\boldsymbol{\rho}^\star}=\mathrm{Diag}^{-1}(\boldsymbol{\rho}^\star)$ and $I_{\boldsymbol{\alpha}^\star}$ the component-wise inverse of $\Sigma_{\boldsymbol{\alpha}^\star}$. For any s, t and u in a compact set, such that $t^T\mathbf{1}_g=0$ and $s^T\mathbf{1}_m=0$, we have:*

$$\begin{aligned}\mathcal{L}_c^\star&\left(\boldsymbol{\pi}^\star+\frac{s}{\sqrt{n}},\boldsymbol{\rho}^\star+\frac{t}{\sqrt{d}},\boldsymbol{\alpha}^\star+\frac{u}{\sqrt{nd}}\right)\\&=\mathcal{L}_c^\star\left(\boldsymbol{\theta}^\star\right)+s^T\mathbf{Y}_{\boldsymbol{\pi}^\star}+t^T\mathbf{Y}_{\boldsymbol{\rho}^\star}+\mathrm{Tr}(u^T\mathbf{Y}_{\boldsymbol{\alpha}^\star})\\&-\left(\frac{1}{2}s^TI_{\boldsymbol{\pi}^\star}s+\frac{1}{2}t^TI_{\boldsymbol{\rho}^\star}t+\frac{1}{2}\mathrm{Tr}[(u\odot u)^TI_{\boldsymbol{\alpha}^\star}]\right)\\&+o_P(1)\end{aligned}$$

where $\odot$ denotes the Hadamard product of two matrices (element-wise product), $\mathbf{Y}_{\boldsymbol{\pi}^\star}$, $\mathbf{Y}_{\boldsymbol{\rho}^\star}$ are asymptotically centered Gaussian vectors of sizes g and m with respective variance matrices $I(\boldsymbol{\pi}^\star)$ and $I(\boldsymbol{\rho}^\star)$ and $\mathbf{Y}_{\boldsymbol{\alpha}^\star}$ is a random matrix of size $g\times m$ with independent Gaussian components $Y_{\boldsymbol{\alpha}^\star,kl}\sim\mathcal{N}(0,I_{\boldsymbol{\alpha}^\star,kl})$.

Proof. By Taylor expansion, and with the condition $s^T\mathbf{1}_g=t^T\mathbf{1}_m=0$

$$\begin{aligned}&\mathcal{L}_c^\star\left(\boldsymbol{\pi}^\star+\frac{s}{\sqrt{n}},\boldsymbol{\rho}^\star+\frac{t}{\sqrt{d}},\boldsymbol{\alpha}^\star+\frac{u}{\sqrt{nd}}\right)\\=\;&\mathcal{L}_c^\star\left(\boldsymbol{\theta}^\star\right)+\frac{1}{\sqrt{n}}s^T\nabla\mathcal{L}_{c\boldsymbol{\pi}}^\star\left(\boldsymbol{\theta}^\star\right)+\frac{1}{\sqrt{d}}t^T\nabla\mathcal{L}_{c\boldsymbol{\rho}}^\star\left(\boldsymbol{\theta}^\star\right)+\frac{1}{\sqrt{nd}}\mathrm{Tr}\left(u^T\nabla\mathcal{L}_{c\boldsymbol{\alpha}}^\star\left(\boldsymbol{\theta}^\star\right)\right)\\&+\frac{1}{n}s^T\mathbf{H}_{\boldsymbol{\pi}}\left(\boldsymbol{\theta}^\star\right)s+\frac{1}{d}t^T\mathbf{H}_{\boldsymbol{\rho}}\left(\boldsymbol{\theta}^\star\right)t+\frac{1}{nd}\mathrm{Tr}\left((u\odot u)^T\mathbf{H}_{\boldsymbol{\alpha}}\left(\boldsymbol{\theta}^\star\right)\right)+o_P(1)\end{aligned}$$

where $\nabla\mathcal{L}_{c\boldsymbol{\pi}}^\star\left(\boldsymbol{\theta}^\star\right)$, $\nabla\mathcal{L}_{c\boldsymbol{\rho}}^\star\left(\boldsymbol{\theta}^\star\right)$ and $\nabla\mathcal{L}_{c\boldsymbol{\alpha}}^\star\left(\boldsymbol{\theta}^\star\right)$ denote the respective components of the gradient of $\mathcal{L}_c^\star$ evaluated at $\boldsymbol{\theta}^\star$ and $\mathbf{H}_{\boldsymbol{\pi}}$, $\mathbf{H}_{\boldsymbol{\rho}}$ and $\mathbf{H}_{\boldsymbol{\alpha}}$ denotes the conditional hessian of $\mathcal{L}_c^\star$ evaluated at $\boldsymbol{\theta}^\star$. By inspection, $\mathbf{H}_{\boldsymbol{\pi}}/n$, $\mathbf{H}_{\boldsymbol{\rho}}/d$ and $\mathbf{H}_{\boldsymbol{\alpha}}/nd$ converge in probability to constant matrices and the random vectors $\nabla\mathcal{L}_{c\boldsymbol{\pi}}^\star\left(\boldsymbol{\theta}^\star\right)/\sqrt{n}$, $\nabla\mathcal{L}_{c\boldsymbol{\rho}}^\star\left(\boldsymbol{\theta}^\star\right)/\sqrt{d}$ and $\nabla\mathcal{L}_{c\boldsymbol{\alpha}}^\star\left(\boldsymbol{\theta}^\star\right)/\sqrt{nd}$ converge in distribution to Gaussian vectors by the central limit theorem. □

4. Profile likelihood

Our main result compares the observed likelihood ratio $p(\mathbf{x};\boldsymbol{\theta})/p(\mathbf{x};\boldsymbol{\theta}^\star)$ with the complete likelihood $p(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta})/p(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}^\star)$. To study the behavior of these likelihoods, we shall work conditionally to the true configurations $(\mathbf{z}^\star,\mathbf{w}^\star)$ that have enough observations in each row or column group. We therefore define in section 4.1 so called *regular* configurations and prove that they occur with high probability. We then introduce in section 4.2 conditional and profile log-likelihood ratios and state some of their properties.

4.1. Regular assignments

Definition 4.1 (c-regular assignments)**.** *Let* $\mathbf{z}\in\mathcal{Z}$ *and* $\mathbf{w}\in\mathcal{W}$. *For any* $c>0$, *we say that* $\mathbf{z}$ *and* $\mathbf{w}$ *are* c-regular *if*

$$\min_k z_{+k} \geq cn \quad \textit{and} \quad \min_\ell w_{+\ell} \geq cd.$$

In regular configurations, each row-group for example has $\Omega(n)$ members, where $u_n=\Omega(n)$ if there exists two constant $a,b>0$ such that for n enough large $an\leq u_n\leq bn$. $c/2$-regular assignments, with c defined in Assumption H_1, have high $\mathbb{P}_{\boldsymbol{\theta}^\star}$-probability in the space of all assignments, uniformly over all $\boldsymbol{\theta}^\star\in\boldsymbol{\Theta}$, as stated in Proposition 4.2.

Proposition 4.2. *Define* $\mathcal{Z}_1$ *and* $\mathcal{W}_1$ *as the subsets of* $\mathcal{Z}$ *and* $\mathcal{W}$ *made of* $c/2$*-regular assignments, with c defined in assumption* H_1. *Denote* Ω_1 *the event* $\{(\mathbf{z}^\star,\mathbf{w}^\star)\in\mathcal{Z}_1\times\mathcal{W}_1\}$, *then:*

$$\mathbb{P}_{\boldsymbol{\theta}^\star}\left(\bar{\Omega}_1\right) \leq g\exp\left(-\frac{nc^2}{2}\right) + m\exp\left(-\frac{dc^2}{2}\right).$$

Each z_{+k} is a sum of n i.i.d Bernoulli random variables with parameter $\pi_k\geq\pi_{\min}\geq c$. The proof is straightforward and stems from a simple Hoeffding bound

$$\mathbb{P}_{\boldsymbol{\theta}^\star}\left(z_{+k}\leq n\frac{c}{2}\right)\leq\mathbb{P}_{\boldsymbol{\theta}^\star}\left(z_{+k}\leq n\frac{\pi_k}{2}\right)\leq\exp\left(-2n\left(\frac{\pi_k}{2}\right)^2\right)\leq\exp\left(-\frac{nc^2}{2}\right)$$

and a union bound over g values of k, with similar approach for $w_{+\ell}$.

4.2. Conditional and profile log-likelihoods

Introducing the conditional log-likelihood ratio

$$F_{nd}(\boldsymbol{\theta},\mathbf{z},\mathbf{w}) = \log\frac{p(\mathbf{x}|\mathbf{z},\mathbf{w};\boldsymbol{\theta})}{p(\mathbf{x}|\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}^\star)},$$

the complete likelihood can be written as follows

$$p(\mathbf{x},\mathbf{z},\mathbf{w};\boldsymbol{\theta}) = p(\mathbf{z},\mathbf{w};\boldsymbol{\theta})p(\mathbf{x}|\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}^\star)\exp(F_{nd}(\boldsymbol{\theta},\mathbf{z},\mathbf{w})).$$

The study of F_{nd} will be of crucial importance, as well as its maximum over $\boldsymbol{\Theta}$. After some definitions, we examine some useful properties.

Definition 4.3. *The conditional expectation G of F_{nd} is defined as:*

$$G(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}) = \mathbb{E}_{\boldsymbol{\theta}^\star}\left[\log \frac{p(\mathbf{x}|\mathbf{z}, \mathbf{w}; \boldsymbol{\theta})}{p(\mathbf{x}|\mathbf{z}^\star, \mathbf{w}^\star; \boldsymbol{\theta}^\star)}\middle| \mathbf{z}^\star, \mathbf{w}^\star\right] = \mathbb{E}_{\boldsymbol{\theta}^\star}\left[F_{nd}(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w})|\, \mathbf{z}^\star, \mathbf{w}^\star\right]$$

Moreover, the profile log-likelihood ratio Λ and its expectation $\tilde{\Lambda}$ are defined as:

$$\Lambda(\mathbf{z}, \mathbf{w}) = \max_{\boldsymbol{\theta}} F_{nd}(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w})$$
$$\tilde{\Lambda}(\mathbf{z}, \mathbf{w}) = \max_{\boldsymbol{\theta}} G(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}).$$

Remark 4.4. As F_{nd} and G only depend on $\boldsymbol{\theta}$ through $\boldsymbol{\alpha}$, we will sometimes replace $\boldsymbol{\theta}$ with $\boldsymbol{\alpha}$ in the expressions of F_{nd} and G. Replacing $F_{n,d}$ and G by their profiled version Λ and $\tilde{\Lambda}$ allows us to get rid of the continuous argument $\boldsymbol{\alpha}$ of F_{nd} and to rely instead only on discrete contrasts Λ and $\tilde{\Lambda}$.

Now, Proposition 4.5 characterizes which values of $\boldsymbol{\alpha}$ maximize F_{nd} and G to reach Λ and $\tilde{\Lambda}$. Propositions 4.6 and 4.7 in turn describes properties of G and $\tilde{\Lambda}$ relative to $(\mathbf{z}, \mathbf{w})$.

Proposition 4.5 (maximum of G and $\tilde{\Lambda}$ in θ)**.** *Let $\widehat{\theta}_c = (\widehat{x}_{k\ell}(\mathbf{z}, \mathbf{w}), \widehat{\pi}_k(\mathbf{z}), \widehat{\rho}_\ell(\mathbf{w}))$ be the maximum likelihood estimator of the complete model, as defined in Equation 3.1. Conditionally on $\mathbf{z}^\star, \mathbf{w}^\star$, define the following quantities:*

$$\boldsymbol{S}^\star = (S^\star_{k\ell})_{k\ell} = (\psi'(\alpha^\star_{k\ell}))_{k\ell}$$
$$\bar{x}_{k\ell}(\mathbf{z}, \mathbf{w}) = \mathbb{E}_{\boldsymbol{\theta}^\star}[\widehat{x}_{k\ell}(\mathbf{z}, \mathbf{w})|\mathbf{z}^\star, \mathbf{w}^\star] = \frac{\left[\mathbb{R}_g(\mathbf{z})^T \boldsymbol{S}^\star \mathbb{R}_m(\mathbf{w})\right]_{k\ell}}{\widehat{\pi}_k(\mathbf{z})\widehat{\rho}_\ell(\mathbf{w})} \tag{4.1}$$

with $\bar{x}_{k\ell}(\mathbf{z}, \mathbf{w}) = 0$ for $\mathbf{z}$ and $\mathbf{w}$ such that $\widehat{\pi}_k(\mathbf{z}) = 0$ or $\widehat{\rho}_\ell(\mathbf{w}) = 0$.

Then $F_{nd}(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w})$ (resp. $G(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w})$) is maximum in $\boldsymbol{\alpha}$ for $\boldsymbol{\alpha} = \widehat{\boldsymbol{\alpha}}(\mathbf{z}, \mathbf{w})$ (resp. $\bar{\boldsymbol{\alpha}}(\mathbf{z}, \mathbf{w})$) defined by:

$$\widehat{\alpha}(\mathbf{z}, \mathbf{w})_{k\ell} = (\psi')^{-1}(\widehat{x}_{k\ell}(\mathbf{z}, \mathbf{w})) \quad \textit{and} \quad \bar{\alpha}(\mathbf{z}, \mathbf{w})_{k\ell} = (\psi')^{-1}(\bar{x}_{k\ell}(\mathbf{z}, \mathbf{w})).$$

Hence,

$$\Lambda(\mathbf{z}, \mathbf{w}) = F_{nd}(\widehat{\boldsymbol{\alpha}}(\mathbf{z}, \mathbf{w}), \mathbf{z}, \mathbf{w})$$
$$\tilde{\Lambda}(\mathbf{z}, \mathbf{w}) = G(\bar{\boldsymbol{\alpha}}(\mathbf{z}, \mathbf{w}), \mathbf{z}, \mathbf{w})$$

Note that although $\bar{x}_{k\ell} = \mathbb{E}_{\boldsymbol{\theta}^\star}\left[\widehat{x}_{k\ell}|\, \mathbf{z}^\star, \mathbf{w}^\star\right]$, in general $\bar{\alpha}_{k\ell} \neq \mathbb{E}_{\boldsymbol{\theta}^\star}\left[\widehat{\alpha}_{k\ell}|\, \mathbf{z}^\star, \mathbf{w}^\star\right]$ by non linearity of $(\psi')^{-1}$. Nevertheless, since $(\psi')^{-1}$ is Lipschitz over compact subsets of $\psi'(\mathring{\mathcal{A}})$, with high probability, $|\bar{\alpha}_{k\ell} - \widehat{\alpha}_{k\ell}|$ and $|\widehat{x}_{k\ell} - \bar{x}_{k\ell}|$ are of the same order of magnitude.

Proposition 4.6 (maximum of G and $\tilde{\Lambda}$ in $(\theta, \mathbf{z}, \mathbf{w})$)**.** *Let $\mathrm{KL}(\alpha, \alpha') = \psi'(\alpha)(\alpha - \alpha') + \psi(\alpha') - \psi(\alpha)$ be the Kullback divergence between $\varphi(., \alpha)$ and $\varphi(., \alpha')$ then:*

$$G(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}) = -nd \sum_{k,k'} \sum_{\ell,\ell'} \mathbb{R}_g(\mathbf{z})_{k,k'} \mathbb{R}_m(\mathbf{w})_{\ell,\ell'} \,\mathrm{KL}(\alpha^\star_{k\ell}, \alpha_{k'\ell'}) \leq 0. \tag{4.2}$$

Conditionally on the set Ω_1 of regular assignments and for $n, d > 2/c$,

(i) G *is maximized at* $(\boldsymbol{\alpha}^\star, \mathbf{z}^\star, \mathbf{w}^\star)$ *and its equivalence class.*
(ii) $\tilde{\Lambda}$ *is maximized at* $(\mathbf{z}^\star, \mathbf{w}^\star)$ *and its equivalence class and* $\tilde{\Lambda}(\mathbf{z}^\star, \mathbf{w}^\star) = 0$*.*

Moreover, the maximum of $\tilde{\Lambda}$ in $(\mathbf{z}^\star, \mathbf{w}^\star)$ is well separated, in the sense that there exists a positive gap between $\tilde{\Lambda}(\mathbf{z}^\star, \mathbf{w}^\star)$ and any other $\tilde{\Lambda}(\mathbf{z}, \mathbf{w})$ for $(\mathbf{z}, \mathbf{w})$ in a close neighborhood of $(\mathbf{z}^\star, \mathbf{w}^\star)$, as stated in the following proposition:

Proposition 4.7 (Separability for $\tilde{\Lambda}$)**.** *Conditionally upon* Ω_1*, there exists a positive constant* C *such that for all* $(\mathbf{z}, \mathbf{w}) \in S(\mathbf{z}^\star, \mathbf{w}^\star, C)$*:*

$$\tilde{\Lambda}(\mathbf{z}, \mathbf{w}) \leq -\frac{c\delta(\boldsymbol{\alpha}^\star)}{4} \left(d\|\mathbf{z} - \mathbf{z}^\star\|_{0,\sim} + n\|\mathbf{w} - \mathbf{w}^\star\|_{0,\sim}\right) \tag{4.3}$$

Moreover, there exists a positive constant $B(C)$ *such that for all* $(\mathbf{z}, \mathbf{w}) \notin S(\mathbf{z}^\star, \mathbf{w}^\star, C)$

$$\tilde{\Lambda}(\mathbf{z}, \mathbf{w}) \leq -B(C)\, nd \tag{4.4}$$

The proofs of these propositions are reported in Appendix A. Proof of Proposition 4.5 follows from a straightforward calculation, proof of Proposition 4.6 uses the technical Lemma D.1 to characterize the maximum of G and proof of Proposition 4.7 uses regularity properties of the gradient of $\tilde{\Lambda}$ to control its behavior near its maximum.

5. Main result

Our main result matches the asymptotics of complete and observed likelihoods and is the key to prove the consistency of maximum likelihood and variational estimators. It is set under the assumptions described in section 2.2 and the following asymptotics for the number of rows n and columns d:

$$(H_4): \quad \log(d)/n \to 0 \text{ and } \log(n)/d \to 0.$$

Theorem 5.1 (complete-observed)**.** *Let* $\mathbf{x}$ *be a matrix of* $n \times d$ *observations of a LBM with true parameter* $\boldsymbol{\theta}^\star = (\boldsymbol{\pi}^\star, \boldsymbol{\rho}^\star, \boldsymbol{\alpha}^\star)$ *where the number of row-groups* g *and column-groups* m *are known, which conditional distribution belongs to a regular univariate exponential family. The true random and unobserved assignations for rows and columns are denoted* $\mathbf{z}^\star$ *and* $\mathbf{w}^\star$ *respectively. Define* $\#\operatorname{Sym}(\boldsymbol{\theta})$ *as the number of pairs of permutations* (s, t) *for which* $\boldsymbol{\theta}$ *exhibits symmetry.*

If assumptions H_1 *to* H_4 *are fulfilled, then, the observed likelihood ratio behaves like the complete likelihood ratio, up to a bounded multiplicative factor:*

$$\frac{p(\mathbf{x}; \boldsymbol{\theta})}{p(\mathbf{x}; \boldsymbol{\theta}^\star)} = \frac{\#\operatorname{Sym}(\boldsymbol{\theta})}{\#\operatorname{Sym}(\boldsymbol{\theta}^\star)} \max_{\boldsymbol{\theta}' \sim \boldsymbol{\theta}} \frac{p(\mathbf{x}, \mathbf{z}^\star, \mathbf{w}^\star; \boldsymbol{\theta}')}{p(\mathbf{x}, \mathbf{z}^\star, \mathbf{w}^\star; \boldsymbol{\theta}^\star)} \left(1 + o_P(1)\right) + o_P(1)$$

where both o_P *are uniform over all* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$*.*

The maximum over all $\boldsymbol{\theta}'$ that are equivalent to $\boldsymbol{\theta}$ stems from the fact that because of label-switching, $\boldsymbol{\theta}$ is only identifiable up to its $\sim$-equivalence class from the observed likelihood, whereas it is completely identifiable from the complete likelihood as in this latter case, the labels are known. The terms $\#\,\mathrm{Sym}$ are needed to take into account cases where $\boldsymbol{\theta}$ exhibits symmetry. These were omitted by [2] for SBM, although they are also needed in this case, see remark 5.3. When no $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ exhibits symmetry, the following corollary is immediately deduced:

Corollary 5.2. *If $\boldsymbol{\Theta}$ contains only parameters that do not exhibit symmetry:*

$$\frac{p(\mathbf{x};\boldsymbol{\theta})}{p\left(\mathbf{x};\boldsymbol{\theta}^{\star}\right)} = \max_{\boldsymbol{\theta}'\sim\boldsymbol{\theta}} \frac{p(\mathbf{x},\mathbf{z}^{\star},\mathbf{w}^{\star};\boldsymbol{\theta}')}{p(\mathbf{x},\mathbf{z}^{\star},\mathbf{w}^{\star};\boldsymbol{\theta}^{\star})}\left(1+o_P(1)\right)+o_P(1)$$

where the o_P is uniform over all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

General sketch of the proof The proof relies on the following decomposition of the observed likelihood:

$$\begin{aligned} p(\mathbf{x};\boldsymbol{\theta}) &= \sum_{(\mathbf{z},\mathbf{w})} p(\mathbf{x},\mathbf{z},\mathbf{w};\boldsymbol{\theta}) \\ &= \sum_{(\mathbf{z},\mathbf{w})\sim(\mathbf{z}^{\star},\mathbf{w}^{\star})} p(\mathbf{x},\mathbf{z},\mathbf{w};\boldsymbol{\theta}) + \sum_{(\mathbf{z},\mathbf{w})\nsim(\mathbf{z}^{\star},\mathbf{w}^{\star})} p(\mathbf{x},\mathbf{z},\mathbf{w};\boldsymbol{\theta}). \end{aligned}$$

where the second term shall be proved to be asymptotically negligible. Its control stems from the study of the conditional log-likelihood F_{nd}, see Equation 4.2. In fact, the contribution of configurations that are not equivalent to $(\mathbf{z}^{\star},\mathbf{w}^{\star})$ leads itself to the study of a global control, and a sharper local control of F_{nd}. Hence, the proof relies on the examination of the asymptotic behavior of F_{nd} on three types of configurations that partition $\mathcal{Z}\times\mathcal{W}$:

1. *global control* for assignations $(\mathbf{z},\mathbf{w})$ sufficiently far from $(\mathbf{z}^{\star},\mathbf{w}^{\star})$, *i.e.* such that $\tilde{\Lambda}(\mathbf{z},\mathbf{w})$ is of order $\Omega(-nd)$. Proposition 5.5 gives a large deviation result for $F_{nd}-\tilde{\Lambda}(\mathbf{z},\mathbf{w})$ to prove that F_{nd} is also of order $-\Omega_P(nd)$. A key point will be the use of Proposition C.4, establishing a specific concentration inequality for sub-exponential variables. In turn, those assignments contribute as a $o_P(p(\mathbf{x},\mathbf{z}^{\star},\mathbf{w}^{\star};\boldsymbol{\theta}^{\star}))$ to the sum (Proposition 5.6).
2. *local control*: a small deviation result (Proposition 5.7) is needed to show that the combined contribution of assignments close to but not equivalent to $(\mathbf{z}^{\star},\mathbf{w}^{\star})$ is also a $o_P(p(\mathbf{x},\mathbf{z}^{\star},\mathbf{w}^{\star};\boldsymbol{\theta}^{\star}))$ (Proposition 5.8).
3. *equivalent assignments*: Proposition 5.9 examines which of the remaining assignments, all equivalent to $(\mathbf{z}^{\star},\mathbf{w}^{\star})$, contribute to the sum.

Once these propositions proved, the proof is straightforward, as can be seen below. They are in turn carefully presented and discussed in dedicated subsections as they represents the core arguments and their proofs are themselves postponed to Appendix B for more readability.

Proof. We work conditionally to Ω_1, defined in Proposition 4.2, *i.e.*, the high probability event that $(\mathbf{z}^{\star},\mathbf{w}^{\star})$ is a $c/2$-regular assignment. We choose $(\mathbf{z}^{\star},\mathbf{w}^{\star}) \in$

$\mathcal{Z}_1 \times \mathcal{W}_1$ and a sequence t_{nd} decreasing to 0 but satisfying $t_{nd}^2 \gg \frac{n+d}{nd}$. This is possible when $n \to \infty$ and $d \to \infty$, and for example with Assumption (H_4). We write:

$$\begin{aligned} p(\mathbf{x};\boldsymbol{\theta}) = & \sum_{(\mathbf{z},\mathbf{w})\sim(\mathbf{z}^\star,\mathbf{w}^\star)} p(\mathbf{x},\mathbf{z},\mathbf{w};\boldsymbol{\theta}) \\ & + \sum_{(\mathbf{z},\mathbf{w})\notin S(\mathbf{z}^\star,\mathbf{w}^\star,t_{nd})} p(\mathbf{x},\mathbf{z},\mathbf{w};\boldsymbol{\theta}) + \sum_{\substack{(\mathbf{z},\mathbf{w})\in S(\mathbf{z}^\star,\mathbf{w}^\star,t_{nd}) \\ (\mathbf{z},\mathbf{w})\nsim(\mathbf{z}^\star,\mathbf{w}^\star)}} p(\mathbf{z},\mathbf{w},\mathbf{x};\boldsymbol{\theta}) \end{aligned}$$

According to Proposition 5.6, conditionally to Ω_1 and for n, d large enough that $2\sqrt{2nd}t_{nd} \geq gm$, the contribution of far away assignments is

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{(\mathbf{z},\mathbf{w})\notin S(\mathbf{z}^\star,\mathbf{w}^\star,t_{nd})} p(\mathbf{z},\mathbf{w},\mathbf{x};\boldsymbol{\theta}) = o_P(p(\mathbf{z}^\star,\mathbf{w}^\star,\mathbf{x};\boldsymbol{\theta}^\star)).$$

Using the separability of $\tilde{\Lambda}$ and Assumption (H_4), Proposition 5.8 ensures the existence of $C > 0$ such that:

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{\substack{(\mathbf{z},\mathbf{w})\in S(\mathbf{z}^\star,\mathbf{w}^\star,C) \\ (\mathbf{z},\mathbf{w})\nsim(\mathbf{z}^\star,\mathbf{w}^\star)}} p(\mathbf{z},\mathbf{w},\mathbf{x};\boldsymbol{\theta}) = o_P(p(\mathbf{z}^\star,\mathbf{w}^\star,\mathbf{x};\boldsymbol{\theta}^\star))$$

Since t_{nd} decreases to 0, Proposition 5.8 can be applied for the local configurations belonging to $S(t_{nd})$, for n, d large enough. Therefore the observed likelihood ratio reduces to:

$$\frac{p(\mathbf{x};\boldsymbol{\theta})}{p(\mathbf{x};\boldsymbol{\theta}^\star)} = \frac{\displaystyle\sum_{(\mathbf{z},\mathbf{w})\sim(\mathbf{z}^\star,\mathbf{w}^\star)} p(\mathbf{x},\mathbf{z},\mathbf{w};\boldsymbol{\theta}) + p(\mathbf{x};\mathbf{z}^\star,\mathbf{w}^\star,\boldsymbol{\theta}^\star)o_P(1)}{\displaystyle\sum_{(\mathbf{z},\mathbf{w})\sim(\mathbf{z}^\star,\mathbf{w}^\star)} p(\mathbf{x},\mathbf{z},\mathbf{w};\boldsymbol{\theta}^\star) + p(\mathbf{x};\mathbf{z}^\star,\mathbf{w}^\star,\boldsymbol{\theta}^\star)o_P(1)}$$

Proposition 5.9 deals with equivalence and symmetry and allows us to conclude

$$\frac{p(\mathbf{x};\boldsymbol{\theta})}{p(\mathbf{x};\boldsymbol{\theta}^\star)} = \frac{\#\operatorname{Sym}(\boldsymbol{\theta})}{\#\operatorname{Sym}(\boldsymbol{\theta}^\star)} \max_{\boldsymbol{\theta}'\sim\boldsymbol{\theta}} \frac{p(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}')}{p(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}^\star)}(1 + o_P(1)) + o_P(1). \qquad \square$$

Remark 5.3. As already pointed out, if $\boldsymbol{\theta}$ exhibits symmetry, the maximum likelihood assignment is not unique under $\boldsymbol{\theta}$, and $\#\operatorname{Sym}(\boldsymbol{\theta})$ terms contribute with the same weight. This was not taken into account by [2], and it is interesting to see why it should be also present for SBM. Recall that SBM has only one set of labels $\mathbf{z}$. The proof relies on the the decomposition

$$p(\mathbf{x};\boldsymbol{\theta}) = \sum_{\mathbf{z}} p(\mathbf{x},\mathbf{z};\boldsymbol{\theta}) = \sum_{\mathbf{z}'\sim\mathbf{z}^\star} p(\mathbf{x},\mathbf{z}';\boldsymbol{\theta}) + \sum_{\mathbf{z}'\nsim\mathbf{z}^\star} p(\mathbf{x},\mathbf{z}';\boldsymbol{\theta})$$

where the second term of the sum is neglectible compared to the first term. Now, $\mathbf{z}' \sim \mathbf{z}^\star$ means that there exists a permutation $t : [g] \to [g]$ such that $\mathbf{z}' = \mathbf{z}^t$

and $p(\mathbf{x},\mathbf{z}^t;\boldsymbol{\theta}) = p(\mathbf{x},\mathbf{z};\boldsymbol{\theta}^t)$. The first term is written on Page 1941, Equation (25) in [2] as

$$\sum_{\mathbf{z}'\sim\mathbf{z}^\star} p(\mathbf{x},\mathbf{z}';\boldsymbol{\theta}) = \sum_{\boldsymbol{\theta}'\sim\boldsymbol{\theta}} p(\mathbf{x},\mathbf{z}^\star;\boldsymbol{\theta}') = (1+o(1))\max_{\boldsymbol{\theta}'\sim\boldsymbol{\theta}} p(\mathbf{x},\mathbf{z}^\star;\theta')$$

However, the first equality is not always correct. Actually, we have

$$\sum_{\mathbf{z}'\sim\mathbf{z}^\star} p(\mathbf{x},\mathbf{z}';\boldsymbol{\theta}) = \sum_{t:[g]\to[g]} p(\mathbf{x},\mathbf{z}^{\star,t};\boldsymbol{\theta}) = \sum_{t:[g]\to[g]} p(\mathbf{x},\mathbf{z}^\star;\boldsymbol{\theta}^t)$$

Take a special case of symmetry where $\pi = (1/g,\dots,1/g)$ and $\alpha = (p-q)I_g + q1_g1_g^T$. Then we have $\boldsymbol{\theta}^t = \boldsymbol{\theta}$ for all t. Thus,

$$\sum_{\mathbf{z}'\sim\mathbf{z}^\star} p(\mathbf{x},\mathbf{z}';\boldsymbol{\theta}) = g!\, p(\mathbf{x},\mathbf{z}^\star;\boldsymbol{\theta}).$$

Even for the SBM, we thus have generally:

$$\sum_{\mathbf{z}'\sim\mathbf{z}^\star} p(\mathbf{x},\mathbf{z}';\boldsymbol{\theta}) = (1+o(1))\#\operatorname{Sym}(\boldsymbol{\theta})\max_{\boldsymbol{\theta}'\sim\boldsymbol{\theta}} p(\mathbf{x},\mathbf{z}^\star;\boldsymbol{\theta}')$$

5.1. Global control

A large deviation inequality for configurations $(\mathbf{z},\mathbf{w})$ far from $(\mathbf{z}^\star,\mathbf{w}^\star)$ is build and used to prove that far away configurations make a small contribution to $p(\mathbf{x};\boldsymbol{\theta})$. Since we restricted α in a bounded subset of $\mathring{\mathcal{A}}$, there exists two positive values M_α and κ such that $C_\alpha + (-\kappa,\kappa) \subset [-M_\alpha, M_\alpha] \subset \mathring{\mathcal{A}}$. Moreover, the variance of X_α is bounded away from 0 and $+\infty$:

$$\sup_{\alpha\in[-M_\alpha,M_\alpha]} \mathbb{V}(X_\alpha) = \bar{\sigma}^2 < +\infty \quad \text{and} \quad \inf_{\alpha\in[-M_\alpha,M_\alpha]} \mathbb{V}(X_\alpha) = \underline{\sigma}^2 > 0.$$

Proposition 5.4. *With the previous notations, if $\alpha\in C_\alpha$ and $X_\alpha \sim \varphi(.,\alpha)$, then X_α is sub-exponential with parameters $(\bar{\sigma}^2,\kappa^{-1})$.*

The latter proposition is a direct consequence of the definition of sub-exponential variables, see Appendix C.

Proposition 5.5 (large deviations of F_{nd})**.** *Let* $\operatorname{Diam}(\boldsymbol{\Theta}) = \sup_{\boldsymbol{\theta},\boldsymbol{\theta}'} \|\boldsymbol{\theta}-\boldsymbol{\theta}'\|_\infty$. *For all $\varepsilon_{n,d} \le \kappa\bar{\sigma}$ and n,d*

$$\begin{aligned}&\Delta^1_{nd}(\varepsilon_{nd})\\ &= \mathbb{P}\left(\sup_{\boldsymbol{\theta},\mathbf{z},\mathbf{w}} \left\{F_{nd}(\boldsymbol{\theta},\mathbf{z},\mathbf{w}) - \tilde{\Lambda}(\mathbf{z},\mathbf{w})\right\} \ge \bar{\sigma} nd \operatorname{Diam}(\boldsymbol{\Theta}) 2\sqrt{2}\varepsilon_{nd}\left[1+\frac{gm}{2\sqrt{2nd}\varepsilon_{nd}}\right]\right)\\ &\qquad\qquad \le g^n m^d \exp\left(-\frac{nd\varepsilon^2_{nd}}{2}\right) \qquad (5.1)\end{aligned}$$

In particular, if n and d are large enough that $2\sqrt{2nd}\varepsilon_{nd} \geq gm$, the previous inequality ensures that with high probability, $F_{nd}(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}) - \tilde{\Lambda}(\mathbf{z}, \mathbf{w})$ is no greater than $\bar{\sigma} nd \operatorname{Diam}(\boldsymbol{\Theta}) 4\sqrt{2}\varepsilon_{nd}$.

The concentration inequality used in [2] to prove an analog result for SBM is not sufficient here, as it can be used only for upper-bounded observations, which is obviously not the case for all exponential families. We instead develop a Bernstein-type inequality for sub-exponential variables (Proposition C.4) to upper bound $F_{nd}(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}) - \tilde{\Lambda}(\mathbf{z}, \mathbf{w})$. Proposition 5.5 relies heavily on this Bernstein inequality. A straightforward consequence of this deviation bound is that the combined contribution of assignments far away from $(\mathbf{z}^\star, \mathbf{w}^\star)$ to the sum is negligible, assuming that the numbers n of rows and d of columns grow at commensurate rates, as stated in the following proposition:

Proposition 5.6 (contribution of far away assignments)**.** *Assume $n \to \infty$ and $d \to \infty$, and choose t_{nd} decreasing to 0 such that $t_{nd}^2 \gg \frac{n+d}{nd}$. Then conditionally on Ω_1 and for n, d large enough that $2\sqrt{2nd}t_{nd} \geq gm$, we have:*

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{(\mathbf{z},\mathbf{w}) \notin S(\mathbf{z}^\star, \mathbf{w}^\star, t_{nd})} p(\mathbf{z}, \mathbf{w}, \mathbf{x}; \boldsymbol{\theta}) = p(\mathbf{z}^\star, \mathbf{w}^\star, \mathbf{x}; \boldsymbol{\theta}^\star) o_P(1)$$

where the o_P is uniform in probability over all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

5.2. Local control

Proposition 5.5 gives deviations of order $\mathcal{O}_P(\sqrt{nd})$, which are only useful for $(\mathbf{z}, \mathbf{w})$ such that G and $\tilde{\Lambda}$ are large compared to $\sqrt{nd}$. For $(\mathbf{z}, \mathbf{w})$ close to $(\mathbf{z}^\star, \mathbf{w}^\star)$, we need tighter concentration inequalities, of order $o_P(n + d)$, as follows:

Proposition 5.7 (small deviations F_{nd})**.** *Conditionally upon Ω_1, for n and d satisfying (H_4), and for $(\mathbf{z}, \mathbf{w}) \in S(\mathbf{z}^\star, \mathbf{w}^\star, c/4)$ where $c > 0$ is defined in (H_1), we have:*

$$\sum_{\substack{(\mathbf{z},\mathbf{w}) \in S(\mathbf{z}^\star, \mathbf{w}^\star, c/4) \\ (\mathbf{z},\mathbf{w}) \nsim (\mathbf{z}^\star, \mathbf{w}^\star)}} \frac{\Lambda(\mathbf{z}, \mathbf{w}) - \tilde{\Lambda}(\mathbf{z}^\star, \mathbf{w}^\star)}{d\|\mathbf{z} - \mathbf{z}^\star\|_{0,\sim} + n\|\mathbf{w} - \mathbf{w}^\star\|_{0,\sim}} = o_P(1)$$

The next proposition uses Propositions 4.6 and 5.7 to show that the combined contribution to the observed likelihood of assignments close to $(\mathbf{z}^\star, \mathbf{w}^\star)$ is also a o_P of $p(\mathbf{z}^\star, \mathbf{w}^\star, \mathbf{x}; \boldsymbol{\theta}^\star)$:

Proposition 5.8 (contribution of local assignments)**.** *With the previous notations and for n and d satisfying Assumption (H_4), for any $\tilde{c} \leq \min(C, c/4)$ we have:*

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{\substack{(\mathbf{z},\mathbf{w}) \in S(\mathbf{z}^\star, \mathbf{w}^\star, \tilde{c}) \\ (\mathbf{z},\mathbf{w}) \nsim (\mathbf{z}^\star, \mathbf{w}^\star)}} p(\mathbf{z}, \mathbf{w}, \mathbf{x}; \boldsymbol{\theta}) = o_P(p(\mathbf{z}^\star, \mathbf{w}^\star, \mathbf{x}; \boldsymbol{\theta}^\star))$$

5.3. Equivalent assignments

It remains to study the contribution of equivalent assignments.

Proposition 5.9 (contribution of equivalent assignments)**.** *For all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, we have*

$$\sum_{(\mathbf{z},\mathbf{w})\sim(\mathbf{z}^\star,\mathbf{w}^\star)} \frac{p(\mathbf{x},\mathbf{z},\mathbf{w};\boldsymbol{\theta})}{p(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}^\star)} = \#\,\mathrm{Sym}(\boldsymbol{\theta}) \max_{\boldsymbol{\theta}'\sim\boldsymbol{\theta}} \frac{p(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}')}{p(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}^\star)}(1+o_P(1))$$

where the o_P is uniform in $\boldsymbol{\theta}$.

The maximum over $\boldsymbol{\theta}' \sim \boldsymbol{\theta}$ accounts for equivalent configurations whereas $\#\,\mathrm{Sym}(\boldsymbol{\theta})$ is needed when $\boldsymbol{\theta}$ exhibits symmetry, as noticed in Remark 5.3.

6. Asymptotics for the Maximum Likelihood (MLE) and Variational (VE) Estimators

This section is devoted to the asymptotics of the MLE and VE in the incomplete data model as a consequence of the main result 5.1.

6.1. ML estimator

Theorem 6.1 (Asymptotic behavior of $\widehat{\boldsymbol{\theta}}_{MLE}$)**.** *Denote $\widehat{\boldsymbol{\theta}}_{MLE}$ the maximum likelihood estimator and use the notations of Proposition 3.1. There exist permutations s of $\{1,\dots,g\}$ and t of $\{1,\dots,m\}$ such that*

$$\widehat{\boldsymbol{\pi}}\left(\mathbf{z}^\star\right) - \widehat{\boldsymbol{\pi}}^s_{MLE} = o_P\left(n^{-1/2}\right), \qquad \widehat{\boldsymbol{\rho}}\left(\mathbf{w}^\star\right) - \widehat{\boldsymbol{\rho}}^t_{MLE} = o_P\left(d^{-1/2}\right),$$
$$\widehat{\boldsymbol{\alpha}}\left(\mathbf{z}^\star,\mathbf{w}^\star\right) - \widehat{\boldsymbol{\alpha}}^{s,t}_{MLE} = o_P\left((nd)^{-1/2}\right).$$

The proof relies on a Taylor expansion of the complete likelihood near its optimum, like in Proposition 3.2, and on our main theorem.

Proof. Note first that unless $\boldsymbol{\Theta}$ is constrained and with high probability, $\widehat{\boldsymbol{\theta}}_{MLE}$ and $\widehat{\boldsymbol{\theta}}_c(\mathbf{z}^\star,\mathbf{w}^\star)$ exhibit no symmetries. Indeed, equalities like $\widehat{x}_{k\ell} = \widehat{x}_{k',\ell'}$ have vanishingly small probabilities of being simultaneously true when X_{ij} is discrete and null when X_{ij} is continuous.

Note also that $\mathcal{L}_c^\star$ has a unique maximum at $\widehat{\boldsymbol{\theta}}_c(\mathbf{z}^\star,\mathbf{w}^\star)$. Furthermore, the curvature of $\mathcal{L}_c^\star$ at $\widehat{\boldsymbol{\theta}}_c(\mathbf{z}^\star,\mathbf{w}^\star)$ with respect to $\boldsymbol{\pi}$ (resp. $\boldsymbol{\rho}$, $\boldsymbol{\alpha}$) converge in probability to $I_{\boldsymbol{\pi}^\star}$ (resp. $I_{\boldsymbol{\rho}^\star}$, $I_{\boldsymbol{\alpha}^\star}$) defined in Proposition 3.2 by consistency of $\widehat{\boldsymbol{\theta}}_c$. Therefore any estimator $\widehat{\boldsymbol{\theta}}$ bounded away from $\widehat{\boldsymbol{\theta}}_c(\mathbf{z}^\star,\mathbf{w}^\star)$ satisfies $\mathcal{L}_c^\star(\widehat{\boldsymbol{\theta}}_c(\mathbf{z}^\star,\mathbf{w}^\star)) - \mathcal{L}_c^\star(\widehat{\boldsymbol{\theta}}) > \Omega_P(1)$. If $\|\widehat{\boldsymbol{\theta}}_c(\mathbf{z}^\star,\mathbf{w}^\star) - \widehat{\boldsymbol{\theta}}\| = o_P(1)$, a Taylor expansion at $\widehat{\boldsymbol{\theta}}_c(\mathbf{z}^\star,\mathbf{w}^\star)$ gives

$$\begin{aligned}
&\mathcal{L}_c^\star(\widehat{\boldsymbol{\theta}}_c(\mathbf{z}^\star,\mathbf{w}^\star)) - \mathcal{L}_c^\star(\widehat{\boldsymbol{\theta}}) \\
&\quad = \Big(n(\widehat{\boldsymbol{\pi}}_c\left(\mathbf{z}^\star,\mathbf{w}^\star\right) - \widehat{\boldsymbol{\pi}})^T I_{\boldsymbol{\pi}^\star}(\widehat{\boldsymbol{\pi}}_c\left(\mathbf{z}^\star,\mathbf{w}^\star\right) - \widehat{\boldsymbol{\pi}}) \\
&\qquad + d(\widehat{\boldsymbol{\rho}}_c\left(\mathbf{z}^\star,\mathbf{w}^\star\right) - \widehat{\boldsymbol{\rho}})^T I_{\boldsymbol{\rho}^\star}(\widehat{\boldsymbol{\rho}}_c\left(\mathbf{z}^\star,\mathbf{w}^\star\right) - \widehat{\boldsymbol{\rho}})
\end{aligned}$$

$$+nd\,\mathrm{Tr}\left[\left\{(\widehat{\boldsymbol{\alpha}}_c(\mathbf{z}^\star,\mathbf{w}^\star)-\widehat{\boldsymbol{\alpha}})\odot(\widehat{\boldsymbol{\alpha}}_c(\mathbf{z}^\star,\mathbf{w}^\star)-\widehat{\boldsymbol{\alpha}})\right\}^T I_{\boldsymbol{\alpha}^\star}\right]\Big)$$
$$\times(1+o_P(1))+o_P(1)$$

where the linear term in the expansion vanishes as $\widehat{\boldsymbol{\theta}}_c(\mathbf{z}^\star,\mathbf{w}^\star)$ is the argmax of $\mathcal{L}_c^\star$ and the Hessian of $\mathcal{L}_c^\star$ at $\widehat{\boldsymbol{\theta}}_c(\mathbf{z}^\star,\mathbf{w}^\star)$ were replaced by their limit in probability.

We may now prove the corollary by contradiction. Assume that $\min_s(\widehat{\boldsymbol{\pi}}^s_{MLE}-\widehat{\boldsymbol{\pi}}(\mathbf{w}^\star))\neq o_P\left(n^{-1/2}\right)$, $\min_t(\widehat{\boldsymbol{\rho}}^t_{MLE}-\widehat{\boldsymbol{\rho}}(\mathbf{w}^\star))\neq o_P\left(d^{-1/2}\right)$ or $\min_{s,t}(\widehat{\boldsymbol{\alpha}}^{s,t}_{MLE}-\widehat{\boldsymbol{\alpha}}(\mathbf{z}^\star,\mathbf{w}^\star))\neq o_P\left(nd^{-1/2}\right)$ where s and t are permutations of $\{1,\dots,g\}$ and $\{1,\dots,m\}$. Plugging $\widehat{\boldsymbol{\theta}}^{s,t}_{MLE}$ in the previous expansion shows that:

$$\min_{s,t}\mathcal{L}_c^\star\left(\widehat{\boldsymbol{\theta}}_c(\mathbf{z}^\star,\mathbf{w}^\star)\right)-\mathcal{L}_c^\star\left(\widehat{\boldsymbol{\theta}}^{s,t}_{MLE}\right)=\Omega_P(1). \tag{6.1}$$

But, since $\widehat{\boldsymbol{\theta}}_c(\mathbf{z}^\star,\mathbf{w}^\star)$ and $\widehat{\boldsymbol{\theta}}_{MLE}$ maximize respectively $\frac{p(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta})}{p(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}^\star)}$ and $\frac{p(\mathbf{x};\boldsymbol{\theta})}{p(\mathbf{x};\boldsymbol{\theta}^\star)}$ and have no symmetries, it follows by Theorem 5.1 that

$$\left|\frac{p\left(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\widehat{\boldsymbol{\theta}}_c(\mathbf{z}^\star,\mathbf{w}^\star)\right)}{p(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}^\star)}-\max_{s,t}\frac{p\left(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\widehat{\boldsymbol{\theta}}^{s,t}_{MLE}\right)}{p(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}^\star)}\right|=o_P(1)$$

which contradicts Equation (6.1) and concludes the proof. □

6.2. Variational estimator

Due to the complex dependence structure of the observations, the maximum likelihood estimator of the LBM is not numerically tractable, even with the EM-algorithm. In practice, a variational approximation can be used, see for example [5]: for any joint distribution $\mathbb{Q}\in\mathcal{Q}$ on $\mathcal{Z}\times\mathcal{W}$ a lower bound of $\mathcal{L}(\boldsymbol{\theta})$ is given by

$$\begin{aligned} J(\mathbb{Q},\boldsymbol{\theta}) &= \mathcal{L}(\boldsymbol{\theta})-KL(\mathbb{Q},p(.,.;\boldsymbol{\theta},\mathbf{x}))\\ &= \mathbb{E}_{\mathbb{Q}}\left[\mathcal{L}_c(\mathbf{z},\mathbf{w};\boldsymbol{\theta})\right]+\mathcal{H}(\mathbb{Q}). \end{aligned}$$

where $\mathcal{H}(\mathbb{Q})=-\mathbb{E}_{\mathbb{Q}}[\log(\mathbb{Q})]$. Choose $\mathcal{Q}$ to be the set of factorized distributions, such that for all $(\mathbf{z},\mathbf{w})$

$$\mathbb{Q}(\mathbf{z},\mathbf{w})=\mathbb{Q}(\mathbf{z})\,\mathbb{Q}(\mathbf{w})=\prod_{i,k}\mathbb{Q}(z_{ik}=1)^{z_{ik}}\prod_{j,\ell}\mathbb{Q}(w_{j\ell}=1)^{w_{j\ell}}$$

allows to obtain tractable expressions of $J(\mathbb{Q},\boldsymbol{\theta})$ as a lower bound of the log-likelihood. The variational estimate $\widehat{\boldsymbol{\theta}}_{var}$ of $\boldsymbol{\theta}$ is defined as

$$\widehat{\boldsymbol{\theta}}_{var}\in\underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\mathrm{argmax}}\max_{\mathbb{Q}\in\mathcal{Q}}J(\mathbb{Q},\boldsymbol{\theta}).$$

The following corollary states that $\widehat{\boldsymbol{\theta}}_{var}$ has the same asymptotics as $\widehat{\boldsymbol{\theta}}_{MLE}$ and $\widehat{\boldsymbol{\theta}}_c$.

Theorem 6.2 (Variational estimate)**.** *Under the assumptions of Theorem 5.1 there exist permutations s of $\{1,\ldots,g\}$ and t of $\{1,\ldots,m\}$ such that*

$$\widehat{\boldsymbol{\pi}}\left(\mathbf{z}^{\star}\right)-\widehat{\boldsymbol{\pi}}_{var}^{s}=o_P\left(n^{-1/2}\right),\qquad \widehat{\boldsymbol{\rho}}\left(\mathbf{w}^{\star}\right)-\widehat{\boldsymbol{\rho}}_{var}^{t}=o_P\left(d^{-1/2}\right),$$
$$\widehat{\boldsymbol{\alpha}}\left(\mathbf{z}^{\star},\mathbf{w}^{\star}\right)-\widehat{\boldsymbol{\alpha}}_{var}^{s,t}\quad=\quad o_P\left(\left(nd\right)^{-1/2}\right).$$

Proof. Remark first that for every $\boldsymbol{\theta}$ and for every $(\mathbf{z},\mathbf{w})$,

$$p\left(\mathbf{x},\mathbf{z},\mathbf{w};\boldsymbol{\theta}\right)\leq\exp\left[J\left(\delta_{\mathbf{z}}\times\delta_{\mathbf{w}},\boldsymbol{\theta}\right)\right]\leq\max_{\mathbb{Q}\in\mathcal{Q}}\exp\left[J\left(\mathbb{Q},\boldsymbol{\theta}\right)\right]\leq p\left(\mathbf{x};\boldsymbol{\theta}\right)$$

where $\delta_{\mathbf{z}}$ denotes the dirac mass on $\mathbf{z}$. By dividing by $p\left(\mathbf{x};\boldsymbol{\theta}^{\star}\right)$, we obtain

$$\frac{p\left(\mathbf{x},\mathbf{z},\mathbf{w};\boldsymbol{\theta}\right)}{p\left(\mathbf{x};\boldsymbol{\theta}^{\star}\right)}\leq\frac{\max_{\mathbb{Q}\in\mathcal{Q}}\exp\left[J\left(\mathbb{Q},\boldsymbol{\theta}\right)\right]}{p\left(\mathbf{x};\boldsymbol{\theta}^{\star}\right)}\leq\frac{p\left(\mathbf{x};\boldsymbol{\theta}\right)}{p\left(\mathbf{x};\boldsymbol{\theta}^{\star}\right)}.$$

As this inequality is true for every couple $(\mathbf{z},\mathbf{w})$, we have in particular:

$$\max_{(\mathbf{z},\mathbf{w})\sim(\mathbf{z}^{\star},\mathbf{w}^{\star})}\frac{p\left(\mathbf{x},\mathbf{z},\mathbf{w};\boldsymbol{\theta}\right)}{p\left(\mathbf{x};\boldsymbol{\theta}^{\star}\right)}=\max_{\boldsymbol{\theta}'\sim\boldsymbol{\theta}}\frac{p\left(\mathbf{x},\mathbf{z}^{\star},\mathbf{w}^{\star};\boldsymbol{\theta}'\right)}{p\left(\mathbf{x};\boldsymbol{\theta}^{\star}\right)}\leq\frac{\max_{\mathbb{Q}\in\mathcal{Q}}\exp\left[J\left(\mathbb{Q},\boldsymbol{\theta}\right)\right]}{p\left(\mathbf{x};\boldsymbol{\theta}^{\star}\right)}.$$

Noticing that $p\left(\mathbf{x};\boldsymbol{\theta}^{\star}\right)=\#\operatorname{Sym}(\boldsymbol{\theta}^{\star})p\left(\mathbf{x},\mathbf{z}^{\star},\mathbf{w}^{\star};\boldsymbol{\theta}^{\star}\right)\left(1+o_p(1)\right)$, Theorem 5.1 therefore leads to the following bounds:

$$\begin{aligned}\max_{\boldsymbol{\theta}'\sim\boldsymbol{\theta}}\frac{p\left(\mathbf{x},\mathbf{z}^{\star},\mathbf{w}^{\star};\boldsymbol{\theta}'\right)}{p\left(\mathbf{x},\mathbf{z}^{\star},\mathbf{w}^{\star};\boldsymbol{\theta}^{\star}\right)}(1+o_P(1))&\leq\frac{\max_{\mathbb{Q}\in\mathcal{Q}}\exp\left[J\left(\mathbb{Q},\boldsymbol{\theta}\right)\right]}{p\left(\mathbf{x},\mathbf{z}^{\star},\mathbf{w}^{\star};\boldsymbol{\theta}^{\star}\right)}\\&\leq\#\operatorname{Sym}(\boldsymbol{\theta})\max_{\boldsymbol{\theta}'\sim\boldsymbol{\theta}}\frac{p\left(\mathbf{x},\mathbf{z}^{\star},\mathbf{w}^{\star};\boldsymbol{\theta}'\right)}{p\left(\mathbf{x},\mathbf{z}^{\star},\mathbf{w}^{\star};\boldsymbol{\theta}^{\star}\right)}(1+o_P(1))+o_P(1).\end{aligned}$$

Again, unless $\boldsymbol{\Theta}$ is constrained, $\widehat{\boldsymbol{\theta}}_{VAR}$ exhibits no symmetries with high probability and the same proof by contradiction as in section 6.1 gives the result. □

7. Conclusion

The Latent Block Model offers challenging theoretical questions. We solved under mild assumptions the consistency and asymptotic normality of the maximum likelihood and variational estimators for observations with conditional density belonging to a univariate exponential family, and for a balanced asymptotic rate between the number of rows n and the number of columns d: $\log(d)/n\to0$ and $\log(n)/d\to0$ as n and d tend to infinity. Our results extend those of [2] for binary SBM not only by managing the double direction of LBM, but also by considering larger types of observations. That brought us to define specific concentration inequalities as large and moderate deviations concerning sub-exponential variables. Moreover, we dealt with specific cases of symmetry that were not taken into account as of now.

A specific framework of sparsity was studied by [2]. This is especially convenient for SBM, to model reasonable network settings: increasing sparsity (*i.e.* number of 0) can be done directly by scaling the Bernoulli parameters $p_{k\ell}$ with a common factor that should decrease no faster than $\Omega(\log^{\delta}(n)/n)$, with $\delta > 2$, to ensure consistency. This could also be considered for binary LBM. However this approach fails to model actual observations in the more general valued setting. The equivalent approach could be to consider the product of a Bernoulli variable with the actual observation value. Note however, than even without considering sparsity we recover essentially the same rate: in the sparse-SBM case, each node should be connected to $\Omega(\log^{\delta}(n))$ others to ensure consistency whereas in the dense-LBM case, each of the n-row should should be characterized by $\Omega(\log^{\delta}(n))$ columns (and vice-versa) to ensure consistency.

Alternative research direction could be to explore asymptotic settings where the numbers n of rows and d columns grow at very different rates. Other open question concern estimation of the number of row and column groups and settings where the number of groups increases with n and d.

Appendix A: Proofs of section 4

A.1. Proof of Proposition 4.5 (maximum of G and $\tilde{\Lambda}$ in $\boldsymbol{\theta}$)

Proof. Define $\nu(x,\alpha) = x\alpha - \psi(\alpha)$. For x fixed, $\nu(x,\alpha)$ is maximized at $\alpha = (\psi')^{-1}(x)$. Manipulations yield

$$\begin{aligned} F_{nd}(\boldsymbol{\alpha},\mathbf{z},\mathbf{w}) =& \log p(\mathbf{x};\mathbf{z},\mathbf{w},\boldsymbol{\theta}) - \log p(\mathbf{x};\mathbf{z}^{\star},\mathbf{w}^{\star},\boldsymbol{\theta}^{\star}) \\ =& nd\left[\sum_k\sum_\ell \widehat{\pi}_k(\mathbf{z})\widehat{\rho}_\ell(\mathbf{w})\nu(\widehat{x}_{k\ell}(\mathbf{z},\mathbf{w}),\alpha_{k\ell})\right. \\ &\left. - \sum_k\sum_\ell \widehat{\pi}_k(\mathbf{z}^{\star})\widehat{\rho}_\ell(\mathbf{w}^{\star})\nu(\widehat{x}_{k\ell}(\mathbf{z}^{\star},\mathbf{w}^{\star}),\alpha^{\star}_{k\ell})\right] \end{aligned}$$

which is maximized at $\alpha_{k\ell} = (\psi')^{-1}(\widehat{x}_{k\ell}(\mathbf{z},\mathbf{w}))$. Similarly

$$\begin{aligned} G(\boldsymbol{\alpha},\mathbf{z},\mathbf{w}) =& \mathbb{E}_{\boldsymbol{\theta}^{\star}}\left[\log p(\mathbf{x};\mathbf{z},\mathbf{w},\boldsymbol{\theta}) - \log p(\mathbf{x};\mathbf{z}^{\star},\mathbf{w}^{\star},\boldsymbol{\theta}^{\star})|\mathbf{z}^{\star},\mathbf{w}^{\star}\right] \\ =& nd\left[\sum_k\sum_\ell \widehat{\pi}_k(\mathbf{z})\widehat{\rho}_\ell(\mathbf{w})\nu(\bar{x}_{k\ell}(\mathbf{z},\mathbf{w}),\alpha_{k\ell})\right. \\ &\left. - \sum_k\sum_\ell \widehat{\pi}_k(\mathbf{z}^{\star})\widehat{\rho}_\ell(\mathbf{w}^{\star})\nu(\psi'(\alpha^{\star}_{k\ell}),\alpha^{\star}_{k\ell})\right] \end{aligned}$$

is maximized at $\alpha_{k\ell} = (\psi')^{-1}(\bar{x}_{k\ell}(\mathbf{z},\mathbf{w}))$. □

A.2. Proof of Proposition 4.6 (maximum of G and $\tilde{\Lambda}$ in $(\boldsymbol{\theta},\mathbf{z},\mathbf{w})$)

Proof. We condition on $(\mathbf{z}^{\star},\mathbf{w}^{\star})$ and prove Equation (4.2):

$$\begin{aligned} &G(\boldsymbol{\theta},\mathbf{z},\mathbf{w}) \\ &= \mathbb{E}_{\boldsymbol{\theta}^{\star}}\left[\log\frac{p(\mathbf{x};\mathbf{z},\mathbf{w},\boldsymbol{\theta})}{p(\mathbf{x};\mathbf{z}^{\star},\mathbf{w}^{\star},\boldsymbol{\theta}^{\star})}\middle|\mathbf{z}^{\star},\mathbf{w}^{\star}\right] \end{aligned}$$

$$= \sum_i \sum_j \sum_{k,k'} \sum_{\ell,\ell'} \mathbb{E}_{\boldsymbol{\theta}^\star} \left[x_{ij}(\alpha_{k'\ell'} - \alpha^\star_{k\ell}) - (\psi(\alpha_{k'\ell'}) - \psi(\alpha^\star_{k\ell})) \right] z^\star_{ik} z_{ik'} w^\star_{j\ell} w_{j\ell'}$$

$$= nd \sum_{k,k'} \sum_{\ell,\ell'} \mathbb{R}_g(\mathbf{z})_{k,k'} \mathbb{R}_m(\mathbf{w})_{\ell,\ell'} \left[\psi'(\alpha^\star_{k\ell})(\alpha_{k'\ell'} - \alpha^\star_{k\ell}) + \psi(\alpha^\star_{k\ell}) - \psi(\alpha_{k'\ell'}) \right]$$

$$= -nd \sum_{k,k'} \sum_{\ell,\ell'} \mathbb{R}_g(\mathbf{z})_{k,k'} \mathbb{R}_m(\mathbf{w})_{\ell,\ell'} \, \mathrm{KL}(\alpha^\star_{k\ell}, \alpha_{k'\ell'})$$

If $(\mathbf{z}^\star, \mathbf{w}^\star)$ is regular, and for $n, d > 2/c$, all the rows of $\mathbb{R}_g(\mathbf{z})$ and $\mathbb{R}_m(\mathbf{w})$ have at least one positive element and we can apply lemma D.1 (which is an adaptation for LBM of Lemma 3.2 of [2] for SBM) to characterize the maximum for G.

The maximality of $\tilde{\Lambda}(\mathbf{z}^\star, \mathbf{w}^\star)$ results from the fact that $\tilde{\Lambda}(\mathbf{z}, \mathbf{w}) = G(\bar{\boldsymbol{\alpha}}(\mathbf{z}, \mathbf{w}), \mathbf{z}, \mathbf{w})$ where $\bar{\boldsymbol{\alpha}}(\mathbf{z}, \mathbf{w})$ is a particular value of $\boldsymbol{\alpha}$, $\tilde{\Lambda}$ is immediately maximum at $(\mathbf{z}, \mathbf{w}) \sim (\mathbf{z}^\star, \mathbf{w}^\star)$, and for those, we have $\bar{\boldsymbol{\alpha}}(\mathbf{z}, \mathbf{w}) \sim \boldsymbol{\alpha}^\star$.

The separation and local behavior of G around $(\mathbf{z}^\star, \mathbf{w}^\star)$ is a direct consequence of the proposition 4.7. □

A.3. Proof of Proposition 4.7 (Local upper bound for $\tilde{\boldsymbol{\Lambda}}$)

Proof. We work conditionally on $(\mathbf{z}^\star, \mathbf{w}^\star)$. The principle of the proof relies on the extension of $\tilde{\Lambda}$ to a continuous subspace of $M_g([0,1]) \times M_m([0,1])$, in which confusion matrices are naturally embedded. The regularity assumption allows us to work on a subspace that is bounded away from the borders of $M_g([0,1]) \times M_m([0,1])$. The proof then proceeds by (1) computing the gradient of $\tilde{\Lambda}$ at and around its argmax and (2) using those gradients to control the local behavior of $\tilde{\Lambda}$ around its argmax. The local behavior allows in turn to show that $\tilde{\Lambda}$ is well-separated.

Note that $\tilde{\Lambda}$ only depends on $\mathbf{z}$ and $\mathbf{w}$ through $\mathbb{R}_g(\mathbf{z})$ and $\mathbb{R}_m(\mathbf{w})$. We can therefore extend it to matrices $(U, V) \in \mathcal{U}_c \times \mathcal{V}_c$ where $\mathcal{U}$ is the subset of matrices $\mathcal{M}_g([0,1])$ with each row sum higher than $c/2$ and $\mathcal{V}$ is a similar subset of $\mathcal{M}_m([0,1])$.

$$\tilde{\Lambda}(U, V) = -nd \sum_{k,k'} \sum_{\ell,\ell'} U_{kk'} V_{\ell\ell'} \, \mathrm{KL}\left(\alpha^\star_{k\ell}, \bar{\alpha}_{k'\ell'}\right)$$

where

$$\bar{\alpha}_{k\ell} = \bar{\alpha}_{k\ell}(U, V) = (\psi')^{-1} \left(\frac{\left[U^T \boldsymbol{S}^\star V\right]_{k\ell}}{\left[U^T \mathbf{1} V\right]_{k\ell}} \right)$$

and $\mathbf{1}$ is the $g \times m$ matrix filled with 1. Confusion matrices $\mathbb{R}_g(\mathbf{z})$ and $\mathbb{R}_m(\mathbf{w})$ satisfy $\mathbb{R}_g(\mathbf{z})\mathbb{1} = \widehat{\boldsymbol{\pi}}(\mathbf{z}^\star)$ and $\mathbb{R}_m(\mathbf{w})\mathbb{1} = \widehat{\boldsymbol{\rho}}(\mathbf{w}^\star)$, with $\mathbb{1} = (1, \ldots, 1)^T$ a vector only containing 1 values, and are obviously in $\mathcal{U}_c$ and $\mathcal{V}_c$ as soon as $(\mathbf{z}^\star, \mathbf{w}^\star)$ is $c/2$ regular.

The maps $f_{k,\ell} : (U, V) \mapsto KL(\alpha^\star_{k\ell}, \bar{\alpha}_{k\ell}(U, V))$ are twice differentiable with second derivatives bounded over $\mathcal{U}_c \times \mathcal{V}_c$ and therefore so is $\tilde{\Lambda}(U, V)$. Tedious but straightforward computations show that the derivative of $\tilde{\Lambda}$ at $(D_\pi, D_\rho) \coloneqq (\mathrm{Diag}(\widehat{\boldsymbol{\pi}}(\mathbf{z}^\star)), \mathrm{Diag}(\widehat{\boldsymbol{\rho}}(\mathbf{w}^\star)))$ is:

$$A_{kk'}(\mathbf{w}^\star) \coloneqq -\frac{1}{nd}\frac{\partial \tilde{\Lambda}}{\partial U_{kk'}}(D_\pi, D_\rho) = \sum_\ell \widehat{\rho}_\ell(\mathbf{w}^\star)\,\mathrm{KL}\,(\alpha^\star_{k\ell}, \alpha^\star_{k'\ell})$$
$$B_{\ell\ell'}(\mathbf{z}^\star) \coloneqq -\frac{1}{nd}\frac{\partial \tilde{\Lambda}}{\partial V_{\ell\ell'}}(D_\pi, D_\rho) = \sum_k \widehat{\pi_k}(\mathbf{z}^\star)\,\mathrm{KL}\,(\alpha^\star_{k\ell}, \alpha^\star_{k\ell'})$$

$A(\mathbf{w}^\star)$ and $B(\mathbf{z}^\star)$ are the matrix-derivative of $-\tilde{\Lambda}/nd$ at (D_π, D_ρ). Since $(\mathbf{z}^\star, \mathbf{w}^\star)$ is $c/2$-regular and by definition of $\delta(\boldsymbol{\alpha}^\star)$, $A(\mathbf{w}^\star)_{kk'} \geq c\delta(\boldsymbol{\alpha}^\star)/2$ (resp. $B(\mathbf{w}^\star)_{\ell\ell'} \geq c\delta(\boldsymbol{\alpha}^\star)/2$) if $k \neq k'$ (resp. $\ell \neq \ell'$) and $A(\mathbf{w}^\star)_{kk} = 0$ (resp. $B(\mathbf{z}^\star)_{\ell\ell} = 0$) for all k (resp. ℓ). By boundedness of the second derivative, there exists $C > 0$ such that for all (D_π, D_ρ) and all $(H, G) \in S(D_\pi, D_\rho, C)$, where the definition of the set of local assignments S is extended to the subset of matrices, we have:

$$\frac{-1}{nd}\frac{\partial \tilde{\Lambda}}{\partial U_{kk'}}(H,G)\begin{cases}\geq \frac{3c\delta(\boldsymbol{\alpha}^\star)}{8} & \text{if } k \neq k' \\ \leq \frac{c\delta(\boldsymbol{\alpha}^\star)}{8} & \text{if } k = k'\end{cases}$$
$$\text{and} \quad \frac{-1}{nd}\frac{\partial \tilde{\Lambda}}{\partial V_{\ell\ell'}}(H,G)\begin{cases}\geq \frac{3c\delta(\boldsymbol{\alpha}^\star)}{8} & \text{if } \ell \neq \ell' \\ \leq \frac{c\delta(\boldsymbol{\alpha}^\star)}{8} & \text{if } \ell = \ell'\end{cases}$$

Choose U and V in $(\mathcal{U}_c \times \mathcal{V}_c) \cap S(D_\pi, D_\rho, C)$ satisfying $U\mathbb{1} = \boldsymbol{\pi}(\mathbf{z}^\star)$ and $V\mathbb{1} = \boldsymbol{\rho}(\mathbf{w}^\star)$. $U - D_\pi$ and $V - D_\rho$ have nonnegative off diagonal coefficients and negative diagonal coefficients. Furthermore, the coefficients of U, V, D_π, D_ρ sum up to 1 and $\mathrm{Tr}(D_\pi) = \mathrm{Tr}(D_\rho) = 1$. By Taylor expansion, there exists a couple (H, G) also in $(\mathcal{U}_c \times \mathcal{V}_c) \cap S(D_\pi, D_\rho, C)$ such that

$$\begin{aligned}\frac{-1}{nd}\tilde{\Lambda}\,(U,V) &= \frac{-1}{nd}\left[\tilde{\Lambda}\,(D_\pi, D_\rho) + \mathrm{Tr}\left((U - D_\pi)\frac{\partial \tilde{\Lambda}}{\partial U}(H,G)\right)\right.\\ &\qquad\left. + \mathrm{Tr}\left((V - D_\rho)\frac{\partial \tilde{\Lambda}}{\partial V}(H,G)\right)\right]\\ &\geq \frac{c\delta(\boldsymbol{\alpha}^\star)}{8}\left[3\sum_{k\neq k'}(U - D_\pi)_{kk'} + 3\sum_{\ell\neq\ell'}(V - D_\rho)_{\ell\ell'}\right.\\ &\qquad\left. + \sum_k (U - D_\pi)_{kk} + \sum_\ell (V - D_\rho)_{\ell\ell}\right]\\ &= \frac{c\delta(\boldsymbol{\alpha}^\star)}{4}[(1 - \mathrm{Tr}(U)) + (1 - \mathrm{Tr}(V))]\end{aligned}$$

To conclude the proof, assume without loss of generality that $(\mathbf{z}, \mathbf{w}) \in S(\mathbf{z}^\star, \mathbf{w}^\star, C)$ achieves the $\|.\|_{0,\sim}$ norm (i.e. it is the closest to $(\mathbf{z}^\star, \mathbf{w}^\star)$ in its representative class). Then $(U, V) = (\mathbb{R}_g(\mathbf{z}), \mathbb{R}_m(\mathbf{w}))$ is in $(\mathcal{U}_c \times \mathcal{V}_c) \cap S(D_\pi, D_\rho, C)$ and satisfy $U\mathbb{1} = \boldsymbol{\pi}(\mathbf{z}^\star)$ (resp. $V\mathbb{1} = \boldsymbol{\rho}(\mathbf{w}^\star)$). We just need to note $n(1 -$

$\mathrm{Tr}(\mathbb{R}_g(\mathbf{z}))) = \|\mathbf{z}-\mathbf{z}^\star\|_{0,\sim}$ (resp. $d(1-\mathrm{Tr}(\mathbb{R}_m(\mathbf{w}))) = \|\mathbf{w}-\mathbf{w}^\star\|_{0,\sim}$) to end the proof. □

Appendix B: Proofs of section 5

B.1. Proof of Proposition 5.5 (large deviation for F_{nd})

Proof. Conditionally upon $(\mathbf{z}^\star, \mathbf{w}^\star)$,

$$\begin{aligned}
F_{nd}(\boldsymbol{\theta},\mathbf{z},\mathbf{w}) - \tilde{\Lambda}(\mathbf{z},\mathbf{w}) &\le F_{nd}(\boldsymbol{\theta},\mathbf{z},\mathbf{w}) - G(\boldsymbol{\theta},\mathbf{z},\mathbf{w}) \\
&= \sum_i \sum_j (\alpha_{z_i w_j} - \alpha^\star_{z_i^\star w_j^\star}) \left(x_{ij} - \psi'(\alpha^\star_{z_i^\star w_j^\star}) \right) \\
&= \sum_{kk'} \sum_{\ell\ell'} (\alpha_{k'\ell'} - \alpha^\star_{k\ell}) \, W_{kk'\ell\ell'} \\
&\le \sup_{\substack{\Gamma \in \mathbb{R}^{g^2 \times m^2} \\ \|\Gamma\|_\infty \le \mathrm{Diam}(\boldsymbol{\Theta})}} \sum_{kk'} \sum_{\ell\ell'} \Gamma_{kk'\ell\ell'} W_{kk'\ell\ell'} := Z
\end{aligned}$$

uniformly in $\boldsymbol{\theta}$, where the $W_{kk'\ell\ell'}$ are independent and defined by:

$$W_{kk'\ell\ell'} = \sum_i \sum_j z^\star_{ik} w^\star_{j\ell} z_{i,k'} w_{j\ell'} \left(x_{ij} - \psi'(\alpha^\star_{k\ell}) \right)$$

is the sum of $nd\mathbb{R}_g(\mathbf{z})_{kk'}\mathbb{R}_m(\mathbf{w})_{\ell\ell'}$ sub-exponential variables with parameters $(\bar{\sigma}^2, 1/\kappa)$ and is therefore itself sub-exponential with parameters $(nd\mathbb{R}_g(\mathbf{z})_{kk'}\mathbb{R}_m(\mathbf{w})_{\ell\ell'}\bar{\sigma}^2, 1/\kappa)$. According to Proposition C.4, $\mathbb{E}_{\boldsymbol{\theta}^\star}[Z|\mathbf{z}^\star,\mathbf{w}^\star] \le gm\,\mathrm{Diam}(\boldsymbol{\Theta})\sqrt{nd\bar{\sigma}^2}$ and Z is sub-exponential with parameters $(nd\,\mathrm{Diam}(\boldsymbol{\Theta})^2(2\sqrt{2})^2\bar{\sigma}^2, 2\sqrt{2}\,\mathrm{Diam}(\boldsymbol{\Theta})/\kappa)$. In particular, for all $\varepsilon_{n,d} < \bar{\sigma}\kappa$

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{\theta}^\star}\left(Z \ge \bar{\sigma} gm\,\mathrm{Diam}(\boldsymbol{\Theta})\sqrt{nd}\left\{ 1 + \frac{\sqrt{8nd}\varepsilon_{n,d}}{gm} \right\} \middle| \mathbf{z}^\star, \mathbf{w}^\star \right) \\
&\qquad \le \mathbb{P}_{\boldsymbol{\theta}^\star}\left(Z \ge \mathbb{E}_{\boldsymbol{\theta}^\star}[Z|\mathbf{z}^\star,\mathbf{w}^\star] + \bar{\sigma}\,\mathrm{Diam}(\boldsymbol{\Theta}) nd 2\sqrt{2}\varepsilon_{n,d} \middle| \mathbf{z}^\star, \mathbf{w}^\star \right) \\
&\qquad\qquad \le \exp\left(-\frac{nd\varepsilon_{n,d}^2}{2} \right)
\end{aligned}$$

We can then remove the conditioning and take a union bound to prove Equation (5.1). □

B.2. Proof of Proposition 5.6 (contribution of far away assignments)

Proof. Conditionally on $(\mathbf{z}^\star, \mathbf{w}^\star)$, we know from proposition 4.6 that $\tilde{\Lambda}$ is maximal in $(\mathbf{z}^\star, \mathbf{w}^\star)$ and its equivalence class. Choose $0 < t_{nd}$ decreasing to 0 but satisfying $t_{nd}^2 \gg \frac{n+d}{nd}$. This is possible as $n \to 0$ and $d \to 0$.

According to equation 4.4, for all $(\mathbf{z},\mathbf{w}) \notin S(\mathbf{z}^\star,\mathbf{w}^\star,C)$

$$\tilde{\Lambda}(\mathbf{z},\mathbf{w}) \leq -B(C)\, nd.$$

Now, according to Equation 4.3, for all $(\mathbf{z},\mathbf{w}) \in S(\mathbf{z}^\star,\mathbf{w}^\star,C) \setminus S(\mathbf{z}^\star,\mathbf{w}^\star,t_{nd})$

$$\tilde{\Lambda}(\mathbf{z},\mathbf{w}) \leq -\frac{c\delta(\boldsymbol{\alpha}^\star)}{4}(n\|\mathbf{w}-\mathbf{w}^\star\|_{0,\sim} + d\|\mathbf{z}-\mathbf{z}^\star\|_{0,\sim}) \leq -\frac{c\delta(\boldsymbol{\alpha}^\star)}{4} ndt_{nd}$$

since either $\|\mathbf{z}-\mathbf{z}^\star\|_{0,\sim} \geq nt_{nd}$ or $\|\mathbf{w}-\mathbf{w}^\star\|_{0,\sim} \geq dt_{nd}$. Hence, for n and d large enough for all $(\mathbf{z},\mathbf{w}) \notin S(\mathbf{z}^\star,\mathbf{w}^\star,t_{nd})$

$$\tilde{\Lambda}(\mathbf{z},\mathbf{w}) \leq -\frac{c\delta(\boldsymbol{\alpha}^\star)}{4} ndt_{nd} \tag{B.1}$$

Set $\varepsilon_{nd} = \inf\left(\frac{c\delta(\boldsymbol{\alpha}^\star)t_{nd}}{32\sqrt{2}\bar{\sigma}\,\mathrm{Diam}(\boldsymbol{\Theta})}, \kappa\bar{\sigma}\right)$. By proposition 5.5, and with our choice of ε_{nd}, with probability higher than $1-\Delta^1_{nd}(\varepsilon_{nd})$,

$$\begin{aligned}
&\sum_{(\mathbf{z},\mathbf{w})\notin S(\mathbf{z}^\star,\mathbf{w}^\star,t_{nd})} p(\mathbf{x},\mathbf{z},\mathbf{w};\boldsymbol{\theta}) \\
&= p(\mathbf{x}|\mathbf{z}^\star,\mathbf{w}^\star,\boldsymbol{\theta}^\star) \sum_{(\mathbf{z},\mathbf{w})\notin S(\mathbf{z}^\star,\mathbf{w}^\star,t_{nd})} p(\mathbf{z},\mathbf{w};\boldsymbol{\theta}) e^{F_{nd}(\boldsymbol{\theta},\mathbf{z},\mathbf{w})-\tilde{\Lambda}(\mathbf{z},\mathbf{w})+\tilde{\Lambda}(\mathbf{z},\mathbf{w})} \\
&\leq p(\mathbf{x}|\mathbf{z}^\star,\mathbf{w}^\star,\boldsymbol{\theta}^\star) \sum_{(\mathbf{z},\mathbf{w})\in \mathcal{Z}\times\mathcal{W}} p(\mathbf{z},\mathbf{w};\boldsymbol{\theta}) e^{F_{nd}(\boldsymbol{\theta},\mathbf{z},\mathbf{w})-\tilde{\Lambda}(\mathbf{z},\mathbf{w})-ndt_{nd}c\delta(\boldsymbol{\alpha}^\star)/4} \\
&\leq p(\mathbf{x}|\mathbf{z}^\star,\mathbf{w}^\star,\boldsymbol{\theta}^\star) \sum_{\mathbf{z},\mathbf{w}} p(\mathbf{z},\mathbf{w};\boldsymbol{\theta}) e^{-ndt_{nd}c\delta(\boldsymbol{\alpha}^\star)/8} \\
&= \frac{p(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}^\star)}{p(\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}^\star)} e^{-ndt_{nd}c\delta(\boldsymbol{\alpha}^\star)/8} \\
&\leq p(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}^\star) \exp\left(-ndt_{nd}\frac{c\delta(\boldsymbol{\alpha}^\star)}{8} + (n+d)\log\frac{1}{c}\right) \\
&= p(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}^\star) o(1)
\end{aligned}$$

where the second line comes from inequality (B.1), the third from the global control studied in Proposition 5.5 and the definition of ε_{nd}, the fourth from the definition of $p(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}^\star)$, the fifth from the bounds on $\boldsymbol{\pi}^\star$ and $\boldsymbol{\rho}^\star$ and the last from $t_{nd} \gg (n+d)/nd$.

In addition, we have $\varepsilon^2_{nd} \gg \frac{n+d}{nd}$ so that $\Delta^1_{nd}(\varepsilon_{nd})$ vanishes and:

$$\sum_{(\mathbf{z},\mathbf{w})\notin S(\mathbf{z}^\star,\mathbf{w}^\star,t_{nd})} p(\mathbf{x},\mathbf{z},\mathbf{w};\boldsymbol{\theta}) = p(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star,\boldsymbol{\theta}^\star) o_P(1)$$

where the o_P is uniform in probability over all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. □

B.3. Proof of Proposition 5.7 (local convergence F_{nd})

Proof. We work conditionally on $(\mathbf{z}^\star, \mathbf{w}^\star) \in \mathcal{Z}_1 \times \mathcal{W}_1$. We assume with no loss of generality that $(\mathbf{z}, \mathbf{w})$ is the representative of its class closest to $(\mathbf{z}^\star, \mathbf{w}^\star)$. Choose $\varepsilon \leq \kappa \underline{\sigma}^2$ small. Manipulation of Λ and $\tilde{\Lambda}$ yield

$$\begin{aligned}\frac{F_{nd}(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}) - \tilde{\Lambda}(\mathbf{z}, \mathbf{w})}{nd} \leq & \frac{\Lambda(\mathbf{z}, \mathbf{w}) - \tilde{\Lambda}(\mathbf{z}, \mathbf{w})}{nd} \\ = & \sum_k \sum_\ell \widehat{\pi}_k(\mathbf{z}) \widehat{\rho}_\ell(\mathbf{w}) \left[f(\widehat{x}_{k\ell}) - f(\bar{x}_{k\ell}) \right] \\ & - \sum_k \sum_\ell \widehat{\pi}_k(\mathbf{z}^\star) \widehat{\rho}_\ell(\mathbf{w}^\star) \alpha^\star_{k\ell} (\widehat{x}^\star_{k\ell} - \bar{x}^\star_{k\ell})\end{aligned}$$

where $f(x) = x(\psi')^{-1}(x) - \psi \circ (\psi')^{-1}(x)$, $\widehat{x}^\star_{k\ell} = \widehat{x}_{k\ell}(\mathbf{z}^\star, \mathbf{w}^\star)$ and $\bar{x}^\star_{k\ell} = \psi'(\alpha^\star_{k\ell})$. The function f is twice differentiable on $\mathring{\mathcal{A}}$ with $f'(x) = -(\psi')^{-1}(x)$ and $f''(x) = -1/\psi'' \circ (\psi')^{-1}(x)$. f' (resp. f'') are bounded over $I = \psi'([-M_\alpha, M_\alpha])$ by M_α (resp. $1/\underline{\sigma}^2$).

Assignments $(\mathbf{z}, \mathbf{w})$ belonging to $S(\mathbf{z}^\star, \mathbf{w}^\star, c/4)$ are also $c/4$-regular. According to Proposition C.2, $\widehat{x}_{k\ell}$ and $\bar{x}_{k\ell}$ are at distance at most ε with probability higher than $1 - 2\exp\left(-\frac{ndc^2\varepsilon^2}{32(\bar{\sigma}^2 + \kappa^{-1}\varepsilon)}\right)$, so that:

$$f(\widehat{x}_{k\ell}) - f(\bar{x}_{k\ell}) = f'(\bar{x}_{k\ell})\left(\widehat{x}_{k\ell} - \bar{x}_{k\ell}\right) + \Omega\left((\widehat{x}_{k\ell} - \bar{x}_{k\ell})^2\right)$$

By Proposition C.2, $(\widehat{x}_{k\ell} - \bar{x}_{k\ell})^2 = \mathcal{O}_P(1/nd)$ where the $\mathcal{O}_P$ is uniform in $\mathbf{z}, \mathbf{w}$ and does not depend on $\mathbf{z}^\star, \mathbf{w}^\star$. Similarly,

$$f'(\bar{x}_{k\ell}) = f'(\bar{x}^\star_{k\ell}) + \Omega(\bar{x}_{k\ell} - \bar{x}^\star_{k\ell}) = \alpha^\star_{k\ell} + \Omega(\bar{x}_{k\ell} - \bar{x}^\star_{k\ell})$$

$\bar{x}_{k\ell}$ is a convex combination of the $S^\star_{k\ell} = \psi'(\alpha^\star_{k\ell})$ therefore,

$$\begin{aligned}|\bar{x}_{k\ell} - \bar{x}^\star_{k\ell}| &= \left| \frac{\left[\mathbb{R}_g(\mathbf{z})^T \boldsymbol{S}^\star \mathbb{R}_m(\mathbf{w})\right]_{k\ell}}{\widehat{\pi}_k(\mathbf{z})\widehat{\rho}_\ell(\mathbf{w})} - \bar{x}^\star_{k\ell} \right| \\ &\leq \left(1 - \frac{\mathbb{R}_g(\mathbf{z})_{kk}\mathbb{R}_m(\mathbf{w})_{\ell\ell}}{\widehat{\pi}_k(\mathbf{z})\widehat{\rho}_\ell(\mathbf{w})}\right)(S^\star_{\max} - S^\star_{\min})\end{aligned}$$

Note that:

$$\begin{aligned}\sum_{k,\ell} \widehat{\pi}_k(\mathbf{z})\widehat{\rho}_\ell(\mathbf{w}) \left(1 - \frac{\mathbb{R}_g(\mathbf{z})_{kk}\mathbb{R}_m(\mathbf{w})_{\ell\ell}}{\widehat{\pi}_k(\mathbf{z})\widehat{\rho}_\ell(\mathbf{w})}\right) &= 1 - \operatorname{Tr}(\mathbb{R}_g(\mathbf{z}))\operatorname{Tr}(\mathbb{R}_m(\mathbf{w})) \\ &\leq \frac{\|\mathbf{z} - \mathbf{z}^\star\|_{0,\sim}}{n} + \frac{\|\mathbf{w} - \mathbf{w}^\star\|_{0,\sim}}{d}\end{aligned}$$

and $\widehat{x}_{k\ell} - \bar{x}_{k\ell} = o_P(1)$. Therefore

$$\sum_{k,\ell} \widehat{\pi}_k(\mathbf{z})\widehat{\rho}_\ell(\mathbf{w})\Omega(\bar{x}_{k\ell} - \bar{x}^\star_{k\ell}) \times (\widehat{x}_{k\ell} - \bar{x}_{k\ell}) = o_P\left(\frac{\|\mathbf{z} - \mathbf{z}^\star\|_{0,\sim}}{n} + \frac{\|\mathbf{w} - \mathbf{w}^\star\|_{0,\sim}}{d}\right)$$

The remaining term writes

$$\sum_{k,\ell} \alpha^\star_{k\ell} \left[\widehat{\pi}_k(\mathbf{z})\widehat{\rho}_\ell(\mathbf{w})(\widehat{x}_{k\ell} - \bar{x}_{k\ell}) - \widehat{\pi}_k(\mathbf{z}^\star)\widehat{\rho}_\ell(\mathbf{w}^\star)(\widehat{x}^\star_{k\ell} - \bar{x}^\star_{k\ell})\right]$$

According to Proposition C.3, this term is $o_P(\left(\frac{\|\mathbf{z}-\mathbf{z}^\star\|_{0,\sim}}{n} + \frac{\|\mathbf{w}-\mathbf{w}^\star\|_{0,\sim}}{d}\right)$ uniformly in $(\mathbf{z},\mathbf{w})$ and $(\mathbf{z}^\star,\mathbf{w}^\star) \in \Omega_1$ as soon as $ne^{-ad} \to 0$ and $de^{-an} \to 0$ for all $a > 0$, which is true under (H_4). It follows that:

$$\sup_{\substack{(\mathbf{z},\mathbf{w})\nsim(\mathbf{z}^\star,\mathbf{w}^\star)\\(\mathbf{z},\mathbf{w})\in S(\mathbf{z}^\star,\mathbf{w}^\star,c/4)}} \frac{\Lambda(\mathbf{z},\mathbf{w}) - \tilde{\Lambda}(\mathbf{z}^\star,\mathbf{w}^\star)}{nd} = o_P\left(\frac{\|\mathbf{z}-\mathbf{z}^\star\|_{0,\sim}}{n} + \frac{\|\mathbf{w}-\mathbf{w}^\star\|_{0,\sim}}{d}\right)$$

□

B.4. Proof of Proposition 5.8 (contribution of local assignments)

Proof. By Proposition 4.2, it is enough to prove that the sum is small compared to $p(\mathbf{z}^\star,\mathbf{w}^\star,\mathbf{x};\boldsymbol{\theta}^\star)$ on Ω_1. We work conditionally on $(\mathbf{z}^\star,\mathbf{w}^\star) \in \mathcal{Z}_1 \times \mathcal{W}_1$. Choose $(\mathbf{z},\mathbf{w})$ in $S(\mathbf{z}^\star,\mathbf{w}^\star,\tilde{c})$. This set is non empty as soon as $\min(\tilde{c}n,\tilde{c}d) > 1$.

$$\log\left(\frac{p(\mathbf{z},\mathbf{w},\mathbf{x};\boldsymbol{\theta})}{p(\mathbf{z}^\star,\mathbf{w}^\star,\mathbf{x};\boldsymbol{\theta}^\star)}\right) = \log\left(\frac{p(\mathbf{z},\mathbf{w};\boldsymbol{\theta})}{p(\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}^\star)}\right) + F_{nd}(\boldsymbol{\theta},\mathbf{z},\mathbf{w})$$

We can assume without loss of generality that $(\mathbf{z},\mathbf{w})$ is the representative closest to $(\mathbf{z}^\star,\mathbf{w}^\star)$ and denote $r_1 = \|\mathbf{z}-\mathbf{z}^\star\|_0$ and $r_2 = \|\mathbf{w}-\mathbf{w}^\star\|_0$. Then:

$$\begin{aligned}F_{nd}(\boldsymbol{\theta},\mathbf{z},\mathbf{w}) &\le \Lambda(\mathbf{z},\mathbf{w}) - \tilde{\Lambda}(\mathbf{z},\mathbf{w}) + \tilde{\Lambda}(\mathbf{z},\mathbf{w})\\ &\le \Lambda(\mathbf{z},\mathbf{w}) - \tilde{\Lambda}(\mathbf{z},\mathbf{w}) - \frac{c\delta(\boldsymbol{\alpha}^\star)}{4}(dr_1 + nr_2)\\ &\le -\frac{c\delta(\boldsymbol{\alpha}^\star)}{4}(dr_1 + nr_2)(1 + o_P(1))\end{aligned}$$

where the first line comes from the definition of Λ, the second line from Proposition 4.7 and the fact that $\tilde{c} < C$ and the third from Proposition 5.7 and the fact that $\tilde{c} < c/4$. Thanks to corollary D.3, we also know that:

$$\log\left(\frac{p(\mathbf{z},\mathbf{w};\boldsymbol{\theta})}{p(\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}^\star)}\right) \le \mathcal{O}_P(1)\exp\left\{M_{c/4}(r_1+r_2)\right\}$$

There are at most $\binom{n}{r_1}\binom{n}{r_2} g^{r_1} m^{r_2}$ assignments $(\mathbf{z},\mathbf{w})$ at distance r_1 and r_2 of $(\mathbf{z}^\star,\mathbf{w}^\star)$ and each of them has at most $g^g m^m$ equivalent configurations. Therefore,

$$\frac{\sum_{\substack{(\mathbf{z},\mathbf{w})\in S(\mathbf{z}^\star,\mathbf{w}^\star,\tilde{c})\\(\mathbf{z},\mathbf{w})\nsim(\mathbf{z}^\star,\mathbf{w}^\star)}} p(\mathbf{z},\mathbf{w},\mathbf{x};\boldsymbol{\theta})}{p(\mathbf{z}^\star,\mathbf{w}^\star,\mathbf{x};\boldsymbol{\theta}^\star)}$$

$$\le \mathcal{O}_P(1)\sum_{r_1+r_2\ge 1}\binom{n}{r_1}\binom{n}{r_2} g^{g+r_1} m^{m+r_2}$$

$$\times \exp\left((r_1+r_2)M_{c/4} - \frac{c\delta(\boldsymbol{\alpha}^\star)}{4}(dr_1+nr_2)(1+o_P(1))\right)$$
$$= \mathcal{O}_P(1)\left(1+e^{\log g + M_{c/4} - d\frac{c\delta(\boldsymbol{\alpha}^\star)(1+o_P(1))}{4}}\right)^n$$
$$\times \left(1+e^{\log m + M_{c/4} - n\frac{c\delta(\boldsymbol{\alpha}^\star)(1+o_P(1))}{4}}\right)^d - 1$$
$$\leq \mathcal{O}_P(1) a_{nd} \exp(a_{nd})$$

where $a_{nd} = ne^{\log g + M_{c/4} - d\frac{c\delta(\boldsymbol{\alpha}^\star)(1+o_P(1))}{4}} + de^{\log m + M_{c/4} - n\frac{c\delta(\boldsymbol{\alpha}^\star)(1+o_P(1))}{4}} = o_P(1)$ as soon as $n \gg \log d$ and $d \gg \log n$. □

B.5. Proof of Proposition 5.9 (contribution of equivalent assignments)

Proof. Choose (s,t) permutations of $\{1,\dots,g\}$ and $\{1,\dots,m\}$ and assume that $\mathbf{z} = \mathbf{z}^{\star,s}$ and $\mathbf{w} = \mathbf{w}^{\star,t}$. Then $p(\mathbf{x},\mathbf{z},\mathbf{w};\boldsymbol{\theta}) = p(\mathbf{x},\mathbf{z}^{\star,s},\mathbf{w}^{\star,t};\boldsymbol{\theta}) = p(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}^{s,t})$. If furthermore $(s,t) \in \mathrm{Sym}(\boldsymbol{\theta})$, $\boldsymbol{\theta}^{s,t} = \boldsymbol{\theta}$ and immediately $p(\mathbf{x},\mathbf{z},\mathbf{w};\boldsymbol{\theta}) = p(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta})$. We can therefore partition the sum as

$$\begin{aligned}\sum_{(\mathbf{z},\mathbf{w})\sim(\mathbf{z}^\star,\mathbf{w}^\star)} p(\mathbf{x},\mathbf{z},\mathbf{w};\boldsymbol{\theta}) &= \sum_{s,t} p(\mathbf{x},\mathbf{z}^{\star,s},\mathbf{w}^{\star,t};\boldsymbol{\theta}) \\ &= \sum_{s,t} p(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}^{s,t}) \\ &= \sum_{\boldsymbol{\theta}'\sim\boldsymbol{\theta}} \#\,\mathrm{Sym}(\boldsymbol{\theta}') p(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}') \\ &= \#\,\mathrm{Sym}(\boldsymbol{\theta}) \sum_{\boldsymbol{\theta}'\sim\boldsymbol{\theta}} p(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}')\end{aligned}$$

The complete likelihood $p(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta})$ is a unimodal function of $\boldsymbol{\theta}$ with mode located in $\widehat{\boldsymbol{\theta}}_c$. By consistency of $\widehat{\boldsymbol{\theta}}_c$, either $p(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}) = o_P(p(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}^\star))$ or $p(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}) = \mathcal{O}_P(p(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}^\star))$ when $\boldsymbol{\theta}$ is in a close neighborhood of $\boldsymbol{\theta}^\star$. In the latter case, any $\boldsymbol{\theta}' \sim \boldsymbol{\theta}$ other than $\boldsymbol{\theta}$ is bounded away from $\boldsymbol{\theta}^\star$ and thus $p(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}') = o_P(p(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}^\star))$. In summary,

$$\sum_{\boldsymbol{\theta}'\sim\boldsymbol{\theta}} \frac{p(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}')}{p(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}^\star)} = \max_{\boldsymbol{\theta}'\sim\boldsymbol{\theta}} \frac{p(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}')}{p(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}^\star)}(1+o_P(1))$$ □

Appendix C: Concentration for sub-exponential variables

Concentration inequalities for sub-exponential variables play a key role: in particular Proposition C.4 for global convergence and Propositions C.2 and C.3 for local convergence. We present here some properties of sub-exponential variables [14], then derives the needed concentration inequalities.

Recall first that a random variable X is sub-exponential with parameters (τ^2, b) if for all λ such that $|\lambda| \leq 1/b$,

$$\mathbb{E}[e^{\lambda(X-\mathbb{E}(X))}] \leq \exp\left(\frac{\lambda^2\tau^2}{2}\right).$$

In particular, all distributions coming from a natural exponential family are sub-exponential. Sub-exponential variables satisfy a large deviation Bernstein-type inequality:

$$\mathbb{P}(X - \mathbb{E}[X] \geq t) \leq \begin{cases} \exp\left(-\frac{t^2}{2\tau^2}\right) & \text{if} \quad 0 \leq t \leq \frac{\tau^2}{b} \\ \exp\left(-\frac{t}{2b}\right) & \text{if} \quad t \geq \frac{\tau^2}{b} \end{cases}$$

So that

$$\mathbb{P}(X - \mathbb{E}[X] \geq t) \leq \exp\left(-\frac{t^2}{2(\tau^2 + bt)}\right)$$

C.1. Properties

The sub-exponential property is preserved by summation and multiplication.

- If X is sub-exponential with parameters (τ^2, b) and $\alpha \in \mathbb{R}$, then so is αX with parameters $(\alpha^2\tau^2, \alpha b)$
- If the X_i, $i = 1, \dots, n$ are sub-exponential with parameters (τ_i^2, b_i) and independent, then so is $X = X_1 + \cdots + X_n$ with parameters $(\sum_i \tau_i^2, \max_i b_i)$

Moreover, Lemma C.1 defines the sub-exponential property of the absolute value of a sub-exponential variable.

Lemma C.1. *If X is a zero mean random variable, sub-exponential with parameters (σ^2, b), then $|X|$ is sub-exponential with parameters $(8\sigma^2, 2\sqrt{2}b)$.*

Proof. Denote $\mu = \mathbb{E}|X|$ and consider $Y = |X| - \mu$. Choose λ such that $|\lambda| < (2\sqrt{2}b)^{-1}$. We need to bound $\mathbb{E}[e^{\lambda Y}]$. Note first that $\mathbb{E}[e^{\lambda Y}] \leq \mathbb{E}[e^{\lambda X}] + \mathbb{E}[e^{-\lambda X}] < +\infty$ is properly defined by sub-exponential property of X and we have

$$\mathbb{E}[e^{\lambda Y}] \leq 1 + \sum_{k=2} \frac{|\lambda|^k \mathbb{E}[|Y|^k]}{k!}$$

where we used the fact that $\mathbb{E}[Y] = 0$. We know bound odd moments of $|\lambda Y|$.

$$\mathbb{E}[|\lambda Y|^{2k+1}] \leq (\mathbb{E}[|\lambda Y|^{2k}]\mathbb{E}[|\lambda Y|^{2k+2}])^{1/2} \leq \frac{1}{2}(\lambda^{2k}\mathbb{E}[Y^{2k}] + \lambda^{2k+2}\mathbb{E}[Y^{2k+2}])$$

where we used first Cauchy-Schwarz and then the arithmetic-geometric mean inequality. The Taylor series expansion can thus be reduced to

$$\mathbb{E}[e^{\lambda Y}] \leq 1 + \left(\frac{1}{2} + \frac{1}{2.3!}\right)\mathbb{E}[Y^2]\lambda^2$$

$$
\begin{aligned}
&+\sum_{k=2}^{+\infty}\left(\frac{1}{(2k)!}+\frac{1}{2}\left[\frac{1}{(2k-1)!}+\frac{1}{(2k+1)!}\right]\right)\lambda^{2k}\mathbb{E}[Y^{2k}] \\
&\leq \sum_{k=0}^{+\infty} 2^k \frac{\lambda^{2k}\mathbb{E}[Y^{2k}]}{(2k)!} \\
&\leq \sum_{k=0}^{+\infty} 2^{3k} \frac{\lambda^{2k}\mathbb{E}[X^{2k}]}{(2k)!} = \cosh\left(2\sqrt{2}\lambda X\right) = \mathbb{E}\left[\frac{e^{2\sqrt{2}\lambda X}+e^{-2\sqrt{2}\lambda X}}{2}\right] \\
&\leq e^{\frac{8\lambda^2\sigma^2}{2}}
\end{aligned}
$$

where we used the well-known inequality $\mathbb{E}[|X-\mathbb{E}[X]|^k] \leq 2^k\mathbb{E}[|X|^k]$ to substitute $2^{2k}\mathbb{E}[X^{2k}]$ to $\mathbb{E}[Y^{2k}]$. □

C.2. Concentration inequalities

Proposition C.2 (Maximum in $(\mathbf{z},\mathbf{w})$)**.** *Let $(\mathbf{z},\mathbf{w})$ be a configuration and $\widehat{x}_{k,\ell}(\mathbf{z},\mathbf{w})$ resp. $\bar{x}_{k\ell}(\mathbf{z},\mathbf{w})$ be defined in Equations (3.1) and (4.1). Under the assumptions of the section 2.2, for all $\varepsilon > 0$*

$$
\mathbb{P}\left(\max_{\mathbf{z},\mathbf{w}}\max_{k,l}\widehat{\pi}_k(\mathbf{z})\widehat{\rho}_\ell(\mathbf{w})|\widehat{x}_{k,\ell}-\bar{x}_{k\ell}|>\varepsilon\right)\leq 2g^{n+1}m^{d+1}\exp\left(-\frac{nd\varepsilon^2}{2(\bar{\sigma}^2+\kappa^{-1}\varepsilon)}\right). \tag{C.1}
$$

Additionally, the suprema over all $c/2$-regular assignments satisfies:

$$
\mathbb{P}\left(\max_{\mathbf{z}\in\mathcal{Z}_1,\mathbf{w}\in\mathcal{W}_1}\max_{k,l}|\widehat{x}_{k,\ell}-\bar{x}_{k\ell}|>\varepsilon\right)\leq 2g^{n+1}m^{d+1}\exp\left(-\frac{ndc^2\varepsilon^2}{8(\bar{\sigma}^2+\kappa^{-1}\varepsilon)}\right). \tag{C.2}
$$

Note that equations C.1 and C.2 remain valid when replacing $c/2$ by any $\tilde{c} < c/2$.

Proof. The random variables X_{ij} are subexponential with parameters $(\bar{\sigma}^2, 1/\kappa)$. Conditionally to $(\mathbf{z}^\star,\mathbf{w}^\star)$, $z_{+k}w_{+\ell}(\widehat{x}_{k,\ell}-\bar{x}_{k\ell})$ is a sum of $z_{+k}w_{+\ell}$ centered subexponential random variables. By Bernstein's inequality [11], we therefore have for all $t > 0$

$$
\mathbb{P}(z_{+k}w_{+\ell}|\widehat{x}_{k,\ell}-\bar{x}_{k\ell}|\geq t)\leq 2\exp\left(-\frac{t^2}{2(z_{+k}w_{+\ell}\bar{\sigma}^2+\kappa^{-1}t)}\right)
$$

In particular, if $t = ndx$,

$$
\begin{aligned}
\mathbb{P}\left(\widehat{\pi}_k(\mathbf{z})\widehat{\rho}_\ell(\mathbf{w})|\widehat{x}_{k,\ell}-\bar{x}_{k\ell}|\geq x\right) &\leq 2\exp\left(-\frac{ndx^2}{2(\widehat{\pi}_k(\mathbf{z})\widehat{\rho}_\ell(\mathbf{w})\bar{\sigma}^2+\kappa^{-1}x)}\right) \\
&\leq 2\exp\left(-\frac{ndx^2}{2(\bar{\sigma}^2+\kappa^{-1}x)}\right)
\end{aligned}
$$

uniformly over $(\mathbf{z},\mathbf{w})$. Equation (C.1) then results from a union bound. Similarly,

$$\begin{aligned}\mathbb{P}\left(|\widehat{x}_{k,\ell}-\bar{x}_{k\ell}|\geq x\right) &= \mathbb{P}\left(\widehat{\pi}_k(\mathbf{z})\widehat{\rho}_\ell(\mathbf{w})|\widehat{x}_{k,\ell}-\bar{x}_{k\ell}|\geq \widehat{\pi}_k(\mathbf{z})\widehat{\rho}_\ell(\mathbf{w})x\right)\\ &\leq 2\exp\left(-\frac{ndx^2\widehat{\pi}_k(\mathbf{z})^2\widehat{\rho}_\ell(\mathbf{w})^2}{2(\widehat{\pi}_k(\mathbf{z})\widehat{\rho}_\ell(\mathbf{w})\bar{\sigma}^2+\kappa^{-1}x\widehat{\pi}_k(\mathbf{z})\widehat{\rho}_\ell(\mathbf{w}))}\right)\\ &\leq 2\exp\left(-\frac{ndc^2x^2}{8(\bar{\sigma}^2+\kappa^{-1}x)}\right)\end{aligned}$$

Where the last inequality comes from the fact that $c/2$-regular assignments satisfy $\widehat{\pi}_k(\mathbf{z})\widehat{\rho}_\ell(\mathbf{w})\geq c^2/4$. Equation (C.2) then results from a union bound over $\mathcal{Z}_1\times\mathcal{W}_1\subset\mathcal{Z}\times\mathcal{W}$. □

Proposition C.3 (Maximum in non equivalent $(\mathbf{z},\mathbf{w})$)**.** *Let $(\bar{\mathbf{z}},\bar{\mathbf{w}})$ be any configuration and $(\mathbf{z},\mathbf{w})$ the $\sim$-equivalent configuration that achieves $\|\mathbf{z}-\mathbf{z}^\star\|_0=\|\bar{\mathbf{z}}-\mathbf{z}^\star\|_{0,\sim}$ and $\|\mathbf{w}-\mathbf{w}^\star\|_0=\|\bar{\mathbf{w}}-\mathbf{w}^\star\|_{0,\sim}$, let $\widehat{x}_{k\ell}=\widehat{x}_{k,\ell}(\mathbf{z},\mathbf{w})$ (resp. $\bar{x}_{k\ell}(\mathbf{z},\mathbf{w})$) and $\widehat{x}^\star_{k\ell}=\widehat{x}_{k,\ell}(\mathbf{z}^\star,\mathbf{w}^\star)$ (resp. $\bar{x}^\star_{k\ell}=\bar{x}_{k\ell}(\mathbf{z}^\star,\mathbf{w}^\star)=\psi'(\alpha^\star_{k\ell})$) be as defined in Equations (3.1) and (4.1). Under the assumptions (H_1) to (H_3), for all $\varepsilon\leq\kappa\bar{\sigma}^2$,*

$$\mathbb{P}\left(\max_{(\bar{\mathbf{z}},\bar{\mathbf{w}})\nsim(\mathbf{z}^\star,\mathbf{w}^\star)}\max_{k,l}\frac{nd\left[\widehat{\pi}_k(\mathbf{z})\widehat{\rho}_\ell(\mathbf{w})(\widehat{x}_{k,\ell}-\bar{x}_{k\ell})-\widehat{\pi}_k(\mathbf{z}^\star)\widehat{\rho}_\ell(\mathbf{w}^\star)(\widehat{x}^\star_{k\ell}-\bar{x}^\star_{k\ell})\right]}{n\|\mathbf{w}-\mathbf{w}^\star\|_0+d\|\mathbf{z}-\mathbf{z}^\star\|_0}>\varepsilon\right)\leq g^{g+1}m^{m+1}a_{nd}e^{a_{nd}}$$

where $a_{nd}=nge^{-\frac{d\varepsilon^2}{\bar{\sigma}^2}}+mde^{-\frac{n\varepsilon^2}{\bar{\sigma}^2}}$.

Proof. Denote $r_1=\|\mathbf{z}-\mathbf{z}^\star\|_0$ and $r_2=\|\mathbf{w}-\mathbf{w}^\star\|_0$. The numerator within the max in the fraction can be expanded to

$$Z_{k\ell}(\mathbf{z},\mathbf{w})=\sum_{i,j}(z_{ik}w_{j\ell}-z^\star_{ik}w^\star_{j\ell})(X_{ij}-\alpha^\star_{z^\star_{ik}w^\star_{j\ell}})$$

and is thus a sum of at most $N=nr_2+dr_1$ non-null centered sub-exponential random variables with parameters $(\bar{\sigma}^2,1/\kappa)$. It is therefore centered sub-exponential with parameters $(N\bar{\sigma}^2,1/\kappa)$. By Bernstein inequality, for all $\varepsilon\leq\kappa\bar{\sigma}^2$ we have

$$\mathbb{P}(Z\geq\varepsilon(nr_2+dr_1))\leq\exp\left(-\frac{(nr_2+dr_1)\varepsilon^2}{2\bar{\sigma}^2}\right).$$

There are at most $\binom{n}{r_1}g^{r_1}g^g$ $\mathbf{z}$ at $\|.\|_{0,\sim}$ distance r_1 of $\mathbf{z}^\star$ and $\binom{d}{r_2}m^{r_2}m^m$ $\mathbf{z}$ at $\|.\|_{0,\sim}$ distance r_2 of $\mathbf{w}^\star$. An union bound shows that:

$$\begin{aligned}&\mathbb{P}\left(\max_{(\bar{\mathbf{z}},\bar{\mathbf{w}})\nsim(\mathbf{z}^\star,\mathbf{w}^\star)}\max_{k,l}\frac{Z_{k\ell}(\mathbf{z},\mathbf{w})}{n\|\mathbf{w}-\mathbf{w}^\star\|_0+d\|\mathbf{z}-\mathbf{z}^\star\|_0}\geq\varepsilon\right)\\ &\quad\leq\sum_{r_1+r_2\geq1}\sum_{\substack{\|\bar{\mathbf{z}}-\mathbf{z}^\star\|_{0,\sim}=r_1\\ \|\bar{\mathbf{w}}-\mathbf{w}^\star\|_{0,\sim}=r_2}}gm\mathbb{P}(Z_{k\ell}(\mathbf{z},\mathbf{w})\geq\varepsilon(nr_2+dr_1))\end{aligned}$$

$$\leq \sum_{r_1+r_2\geq 1} \binom{n}{r_1}\binom{d}{r_2} g^{g+1} m^{m+1} g^{r_1} m^{r_2} \exp\left(-(nr_2+dr_1)\varepsilon^2/2\bar{\sigma}^2\right)$$
$$= g^{g+1}m^{m+1} \sum_{r_1+r_2\geq 1} \left(ge^{-\frac{d\varepsilon^2}{\bar{\sigma}^2}}\right)^{r_1} \left(me^{-\frac{n\varepsilon^2}{\bar{\sigma}^2}}\right)^{r_2} \leq g^{g+1}m^{m+1}a_{nd}e^{a_{nd}}$$

where $a_{nd} = nge^{-\frac{d\varepsilon^2}{\bar{\sigma}^2}} + mde^{-\frac{n\varepsilon^2}{\bar{\sigma}^2}}$. □

Proposition C.4 (concentration for sub-exponential)**.** *Let $X_1, \ldots, X_n$ be independent zero mean random variables, sub-exponential with parameters (σ_i^2, b_i). Denote $V_0^2 = \sum_i \sigma_i^2$ and $b = \max_i b_i$. Then the random variable Z defined by:*

$$Z = \sup_{\substack{\Gamma\in\mathbb{R}^n \\ \|\Gamma\|_\infty \leq M}} \sum_i \Gamma_i X_i$$

is also sub-exponential with parameters $(8M^2V_0^2, 2\sqrt{2}Mb)$. Moreover $\mathbb{E}[Z] \leq MV_0\sqrt{n}$ so that for all $t > 0$,

$$\mathbb{P}(Z - MV_0\sqrt{n} \geq t) \leq \exp\left(-\frac{t^2}{2(8M^2V_0^2 + 2\sqrt{2}Mbt)}\right)$$

Proof. Note first that Z can be simplified to $Z = M\sum_i |X_i|$. We just need to bound bound $\mathbb{E}[Z]$. The rest of the proposition results from the fact that the $|X_i|$ are subexponential $(8\sigma_i^2, 2\sqrt{2}b_i)$ by Lemma C.1 and standard properties of sums of independent rescaled subexponential variables.

$$\mathbb{E}[Z] = \mathbb{E}\left[\sup_{\substack{\Gamma\in\mathbb{R}^n \\ \|\Gamma\|_\infty \leq M}} \sum_i \Gamma_i X_i\right] = \mathbb{E}\left[\sum_i M|X_i|\right] \leq M\sum_i \sqrt{\mathbb{E}[X_i^2]}$$
$$= M\sum_i \sigma_i \leq M\left(\sum_i 1\right)^{1/2}\left(\sum_i \sigma_i^2\right)^{1/2} = MV_0\sqrt{n}$$

using Cauchy-Schwarz. □

Appendix D: Technical lemmas

Lemma D.1 is the working horse for proving Proposition 4.6. Corollary D.3 is needed for Theorem 5.8 and Lemma D.2 is an intermediate result for Corollary D.3.

Lemma D.1.
Let η and $\bar{\eta}$ be two matrices from $M_{g\times m}(\Theta)$ and $f : \Theta \times \Theta \to \mathbb{R}_+$ a positive function, A a (squared) confusion matrix of size g and B a (squared) confusion matrix of size m. We denote $d_{k\ell k'\ell'} = f(\eta_{k\ell}, \bar{\eta}_{k'\ell'})$. Assume that

- *all the rows of η are distinct;*
- *all the columns η are distinct;*

- *$f(x,y)=0 \Leftrightarrow x=y$;*
- *each row of A has a non zero element;*
- *each row of B has a non zero element;*

and denote

$$\Sigma = \sum_{kk'}\sum_{\ell\ell'} A_{kk'}B_{\ell\ell'}d_{k\ell k'\ell'}$$

Then,

$$\Sigma = 0 \Leftrightarrow \begin{cases} A, B \text{ are permutation matrices } s, t \\ \bar{\eta} = \eta^{s,t} \text{ i.e. } \forall (k,\ell), \bar{\eta}_{k\ell} = \eta_{s(k)t(\ell)} \end{cases}$$

Proof. If A and B are the permutation matrices corresponding to the permutations s et t: $A_{ij}=0$ if $i \neq s(j)$ and $B_{ij}=0$ if $i \neq t(j)$. As each row of A contains a non zero element and as $A_{s(k)k} > 0$ (resp. $B_{s(\ell)\ell} > 0$) for all k (resp. ℓ), the following sum Σ reduces to

$$\Sigma = \sum_{kk'}\sum_{\ell\ell'} A_{kk'}B_{\ell\ell'}d_{k\ell k'\ell'} = \sum_{k}\sum_{\ell} A_{s(k)k}B_{t(\ell)\ell}d_{s(k)t(\ell)k\ell}$$

Σ is null and sum of positive components, each component is null. However, all $A_{s(k)k}$ and $B_{t(\ell)\ell}$ are not null, so that for all (k,ℓ), $d_{s(k)t(\ell)k\ell}=0$ and $\bar{\eta}_{k\ell} = \eta_{s(k)t(\ell)}$. Now, if A is not a permutation matrix while $\Sigma = 0$ (the same reasoning holds for B or both). Then A owns a column k that contains two non zero elements, say A_{k_1k} and A_{k_2k}. Let $\ell \in \{1\dots m\}$, there exists by assumption ℓ' such that $B_{\ell\ell'} \neq 0$. As $\Sigma = 0$, both products $A_{k_1k}B_{\ell\ell'}d_{k_1\ell k\ell'}$ and $A_{k_2k}B_{\ell\ell'}d_{k_2\ell k\ell'}$ are zero.

$$\begin{cases} A_{k_1k}B_{\ell\ell'}d_{k_1\ell k\ell'} &= 0 \\ A_{k_2k}B_{\ell\ell'}d_{k_2\ell k\ell'} &= 0 \end{cases} \Leftrightarrow \begin{cases} d_{k_1\ell k\ell'} &= 0 \\ d_{k_2\ell k\ell'} &= 0 \end{cases} \Leftrightarrow \begin{cases} \eta_{k_1\ell} = \bar{\eta}_{k\ell'} \\ \eta_{k_2\ell} = \bar{\eta}_{k\ell'} \end{cases} \Leftrightarrow \eta_{k_1\ell} = \eta_{k_2\ell}$$

The previous equality is true for all ℓ, thus rows k_1 and k_2 of η are identical, and contradict the assumptions. □

Lemma D.2.

Let $\mathcal{Z}_1$ be the subset of $\mathcal{Z}$ of c-regular configurations, as defined in Definition 4.1. Let $\mathbb{S}^g = \{\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_g) \in [0,1]^g : \sum_{k=1}^g \pi_k = 1\}$ be the g-dimensional simplex and denote $\mathbb{S}^g_c = \mathbb{S}^g \cap [c, 1-c]^g$. Then there exists two positive constants M_c and M'_c such that for all $\mathbf{z}$, $\mathbf{z}^\star$ in $\mathcal{Z}_1$ and all $\boldsymbol{\pi} \in \mathbb{S}^g_c$

$$|\log p(\mathbf{z}; \widehat{\boldsymbol{\pi}}(\mathbf{z})) - \log p(\mathbf{z}^\star; \widehat{\boldsymbol{\pi}}(\mathbf{z}^\star))| \quad \leq \quad M_c \|\mathbf{z} - \mathbf{z}^\star\|_0$$

Proof. Consider the entropy map $H : \mathbb{S}^g \to \mathbb{R}$ defined as $H(\boldsymbol{\pi}) = -\sum_{k=1}^g \pi_k \log(\pi_k)$. The gradient ∇H is uniformly bounded by $\frac{M_c}{2} = \log\frac{1-c}{c}$ in $\|.\|_\infty$-norm over $\mathbb{S}^g \cap [c, 1-c]^g$. Therefore, for all $\boldsymbol{\pi}$, $\boldsymbol{\pi}^\star \in \mathbb{S}^g \cap [c, 1-c]^g$, we have

$$|H(\boldsymbol{\pi}) - H(\boldsymbol{\pi}^\star)| \leq \frac{M_c}{2}\|\boldsymbol{\pi} - \boldsymbol{\pi}^\star\|_1$$

To prove the inequality, we remark that $\mathbf{z} \in \mathcal{Z}_1$ translates to $\widehat{\boldsymbol{\pi}}(\mathbf{z}) \in \mathbb{S}^g \cap [c, 1-c]^g$, that $\log p(\mathbf{z}; \widehat{\boldsymbol{\pi}}(\mathbf{z})) - \log p(\mathbf{z}^\star; \widehat{\boldsymbol{\pi}}(\mathbf{z}^\star)) = n[H(\widehat{\boldsymbol{\pi}}(\mathbf{z})) - H(\widehat{\boldsymbol{\pi}}(\mathbf{z}^\star))]$ and finally that $\|\widehat{\boldsymbol{\pi}}(\mathbf{z}) - \widehat{\boldsymbol{\pi}}(\mathbf{z}^\star)\|_1 \leq \frac{2}{n}\|\mathbf{z} - \mathbf{z}^\star\|_0$. □

Corollary D.3. *Let* $\mathbf{z}^\star$ *(resp.* $\mathbf{w}^\star$*) be* $c/2$*-regular and* $\mathbf{z}$ *(resp.* $\mathbf{w}$*) at* $\|.\|_0$*-distance* $c/4$ *of* $\mathbf{z}^\star$ *(resp.* $\mathbf{w}^\star$*). Then, for all* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$

$$\log \frac{p(\mathbf{z}, \mathbf{w}; \boldsymbol{\theta})}{p(\mathbf{z}^\star, \mathbf{w}^\star; \boldsymbol{\theta}^\star)} \leq \mathcal{O}_P(1) \exp\left\{M_{c/4}(\|\mathbf{z} - \mathbf{z}^\star\|_0 + \|\mathbf{w} - \mathbf{w}^\star\|_0)\right\}$$

Proof. Note then that:

$$\begin{aligned}
& \frac{p(\mathbf{z}, \mathbf{w}; \boldsymbol{\theta})}{p(\mathbf{z}^\star, \mathbf{w}^\star; \boldsymbol{\theta}^\star)} \\
= & \quad \frac{p(\mathbf{z}, \mathbf{w}; \boldsymbol{\pi}, \boldsymbol{\rho})}{p(\mathbf{z}^\star, \mathbf{w}^\star; \boldsymbol{\pi}^\star, \boldsymbol{\rho}^\star)} \\
= & \quad \frac{p(\mathbf{z}, \mathbf{w}; \boldsymbol{\pi}, \boldsymbol{\rho})}{p(\mathbf{z}^\star, \mathbf{w}^\star; \widehat{\boldsymbol{\pi}}(\mathbf{z}^\star), \widehat{\boldsymbol{\rho}}(\mathbf{w}^\star))} \frac{p(\mathbf{z}^\star, \mathbf{w}^\star; \widehat{\boldsymbol{\pi}}(\mathbf{z}^\star), \widehat{\boldsymbol{\rho}}(\mathbf{w}^\star))}{p(\mathbf{z}^\star, \mathbf{w}^\star; \boldsymbol{\pi}^\star, \boldsymbol{\rho}^\star)} \\
\leq & \quad \frac{p(\mathbf{z}, \mathbf{w}; \widehat{\boldsymbol{\pi}}(\mathbf{z}), \widehat{\boldsymbol{\rho}}(\mathbf{w}))}{p(\mathbf{z}^\star, \mathbf{w}^\star; \widehat{\boldsymbol{\pi}}(\mathbf{z}^\star), \widehat{\boldsymbol{\rho}}(\mathbf{w}^\star))} \frac{p(\mathbf{z}^\star, \mathbf{w}^\star; \widehat{\boldsymbol{\pi}}(\mathbf{z}^\star), \widehat{\boldsymbol{\rho}}(\mathbf{w}^\star))}{p(\mathbf{z}^\star, \mathbf{w}^\star; \boldsymbol{\pi}^\star, \boldsymbol{\rho}^\star)} \\
\leq & \quad \exp\left\{M_{c/4}(\|\mathbf{z} - \mathbf{z}^\star\|_0 + \|\mathbf{w} - \mathbf{w}^\star\|_0)\right\} \times \frac{p(\mathbf{z}^\star, \mathbf{w}^\star; \widehat{\boldsymbol{\pi}}(\mathbf{z}^\star), \widehat{\boldsymbol{\rho}}(\mathbf{w}^\star))}{p(\mathbf{z}^\star, \mathbf{w}^\star; \boldsymbol{\pi}^\star, \boldsymbol{\rho}^\star)} \\
\leq & \quad \mathcal{O}_P(1) \exp\left\{M_{c/4}(\|\mathbf{z} - \mathbf{z}^\star\|_0 + \|\mathbf{w} - \mathbf{w}^\star\|_0)\right\}
\end{aligned}$$

where the first inequality comes from the definition of $\widehat{\boldsymbol{\pi}}(\mathbf{z})$ and $\widehat{\boldsymbol{\rho}}(\mathbf{w})$ and the second from Lemma D.2 and the fact that $\mathbf{z}^\star$ and $\mathbf{z}$ (resp. $\mathbf{w}^\star$ and $\mathbf{w}$) are $c/4$-regular. Finally, local asymptotic normality of the MLE for multinomial proportions ensures that $\frac{p(\mathbf{z}^\star, \mathbf{w}^\star; \widehat{\boldsymbol{\pi}}(\mathbf{z}^\star), \widehat{\boldsymbol{\rho}}(\mathbf{w}^\star))}{p(\mathbf{z}^\star, \mathbf{w}^\star; \boldsymbol{\pi}^\star, \boldsymbol{\rho}^\star)} = \mathcal{O}_P(1)$. □

References

[1] Christophe Ambroise and Catherine Matias. New consistent and asymptotically normal parameter estimates for random-graph mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):3–35, 2012. MR2885838

[2] Peter Bickel, David Choi, Xiangyu Chang, Hai Zhang, et al. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *The Annals of Statistics*, 41(4):1922–1943, 2013. MR3127853

[3] Peter J. Bickel and Aiyou Chen. A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073, 2009.

[4] Alain Celisse, Jean-Jacques Daudin, Laurent Pierre, et al. Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics*, 6:1847–1899, 2012. MR2988467

[5] Gérard Govaert and Mohamed Nadif. Clustering with block mixture models. *Pattern Recognition*, 36(2):463–473, 2003.

[6] Gérard Govaert and Mohamed Nadif. Block clustering with Bernoulli mixture models: Comparison of different approaches. *Computational Statistics & Data Analysis*, 52(6):3233–3245, 2008. MR2424788

[7] Gérard Govaert and Mohamed Nadif. Latent block model for contingency table. *Communications in Statistics—Theory and Methods*, 39(3):416–425, 2010. MR2745285

[8] Gérard Govaert and Mohamed Nadif. *Co-clustering*. John Wiley & Sons, 2013.

[9] Christine Keribin, Vincent Brault, Gilles Celeux, and Gérard Govaert. Estimation and selection for the latent block model on categorical data. *Statistics and Computing*, 25(6):1201–1216, 2015. MR3401881

[10] Mahendra Mariadassou and Catherine Matias. Convergence of the groups posterior distribution in latent or stochastic block models. *Bernoulli*, 21(1):537–573, 2015. MR3322330

[11] Pascal Massart. *Concentration inequalities and model selection*, volume 6. Springer, 2007. MR2319879

[12] J. G. Shanthikumar and U. Sumita. A central limit theorem for random sums of random variables. *Operations Research Letters*, 3(3):153–155, 1984. MR0761508

[13] Tom A. B. Snijders and Krzysztof Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14(1):75–100, Jan 1997. MR1449742

[14] Martin J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019. MR3967104