*Submitted to Bernoulli*
arXiv: arXiv:0000.0000

# Consistency and Asymptotic Normality of Latent Blocks Model Estimators

VINCENT BRAULT[1,*] CHRISTINE KERIBIN[2,**] and MAHENDRA MARIADASSOU[3,†]

[1] *Univ. Grenoble Alpes, LJK, F-38000 Grenoble, France*
*CNRS, LJK, F-38000 Grenoble, France E-mail:* [*] vincent.brault@univ-grenoble-alpes.fr

[2] *Laboratoire de Mathmatiques d'Orsay, CNRS, and INRIA Saclay Île de France, Université Paris-Sud, Université Paris-Saclay, F-91405 Orsay, France.*
*E-mail:* [**] christine.keribin@math.u-psud.fr

[3] *MaIAGE, INRA, Université Paris-Saclay, 78352 Jouy-en-Josas, France*
*E-mail:* [†] mahendra.mariadassou@inra.fr

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. We show here 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 $+\infty$, (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 to the maximum likelihood estimate under the observed model. Moreover, the variational estimator is also consistent.

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

## 1. Introduction

Coclustering is an unsupervised way to cluster simultaneously the rows and columns of a data matrix, and can be used in numerous applications such as recommendation systems, genomics or text mining. Among the coclustering methods, the Latent Block Model (LBM) is based on the definition of a probabilistic model.

We observe a data matrix $X = (x_{ij})$ with $n$ rows and $d$ columns and we suppose that there exists a row-partition with $g$ row-classes and a column-partition with $m$ column-classes. The row (resp. column) class for each row (resp. column) is unknown and has to be determined. Once determined, rows and columns can be re-ordered according to this coclustering, to let appear blocks that are homogeneous and distinct. This leads to a parsimonious data representation.

LBM can deal with binary ([6]), Gaussian ([10]), categorical ([9]) or count ([7]) data. Due to the complex dependence structure, neither the likelihood, nor the computation of the distribution of the assignments conditionally to the observations (E-step of the EM algorithm), and therefore the maximum likelihood estimator (MLE) are numerically tractable. Estimation can be however performed either with a variational approximation (leading to an approximate value of the MLE), or with a Bayesian approach (VBayes algorithm or Gibbs sampler). Notice that [9] recommend to perform a Gibbs sampler combined with a VBayes algorithm.

Although these estimation methods give satisfactory results, the consistence and asymptotic normality of the MLE are still an open question. Some partial results exist for LBM, and this question has been solved for SBM (Stochastic Block Model), a special case of LBM where the data is a random graphe encoded by its adjacency matrix (rows and columns represents the same units, so that there is only one partition, the same for rows and columns). [4] proved in their Theorem 3 that under the true parameter value, the distribution of the assignments conditionally to the observations of a binary SBM 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 is not established that such an estimator exists except in some particular cases ([1] for example). [11] presented a unified frame for LBM and SBM in case of observations coming from an exponential family, and showed the consistency of the assignment conditional distribution under all parameter value in a neighborhood of the true value. [3] and [2] proved the consistency and asymptotic normality of the MLE for the binary SBM. Bursting with the preceding approaches, they first studied the asymptotic behavior of the MLE in the complete model (observations and assignments) which is very simple to handle; then, they showed that the complete likelihood and the marginal likelihood have similar asymptotic behavior by the use of a Bernstein inequality for bounded observations.

We extend these results to the double asymptotic framework of LBM, following the way of [2], and for observations coming from some exponential family. Moreover, we introduce the concept of model symmetry which was not pointed out by these authors, but is necessary to set the asymptotic behavior. The asymptotic normality of the variational estimator is also settled, and an application to model selection criteria is presented.

The paper is organized as follows. The model, main assumptions and notations are introduced in Section 2, where model symmetry is also discussed. Section 3 establishes the asymptotic normality of the complete likelihood estimator, and section 4 settles three different types of assignment behaviors. Our main result showing that the observed likelihood behaves like the complete likelihood takes place in section 5, and the consistency of MLE and variational estimator is deduced. Technical proofs are gathered in the appendices.

## 2. Model and assumptions

The LBM assumes a block clustering structure of a data matrix $X = (x_{ij})$ with $n$ rows and $d$ columns, as the Cartesian product of a row partition $\mathbf{z}$ by a column partition $\mathbf{w}$. More precisely,

- row assignments (or labels) $\mathbf{z}_i$, $i = 1, \ldots, n$, are independent from column assignments (or labels) $\mathbf{w}_j$, $j = 1, \ldots, d$ : $p(\mathbf{z}, \mathbf{w}) = p(\mathbf{z})p(\mathbf{w})$;
- row labels are independent, with a common multinomial distribution: $\mathbf{z}_i \sim \mathcal{M}(1, \boldsymbol{\pi} = (\pi_1, \ldots, \pi_g))$; in the same way, column labels are i.i.d. multinomial variables: $\mathbf{w}_j \sim \mathcal{M}(1, \boldsymbol{\rho} = (\rho_1, \ldots, \rho_m))$.
- conditionally to row and column assignments $(\mathbf{z}_1, \ldots, \mathbf{z}_n) \times (\mathbf{w}_1, \ldots, \mathbf{w}_d)$, the observed data $X_{ij}$ are independent, and their (conditional) distribution $\varphi(., \alpha)$ belongs to the same parametric family, which parameter $\alpha$ only depends on the given block:

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

where $z_{ik}$ is the indicator variable of whether row $i$ belongs to row-group $k$ and $w_{j\ell}$ is the indicator variable of whether column $j$ belongs to column-group $\ell$.

Hence, the complete parameter set is $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha}) \in \boldsymbol{\Theta}$, with $\boldsymbol{\alpha} = (\alpha_{11}, \ldots, \alpha_{gm})$ and $\boldsymbol{\Theta}$ the parameter space. Figure 1 summarizes these notations.

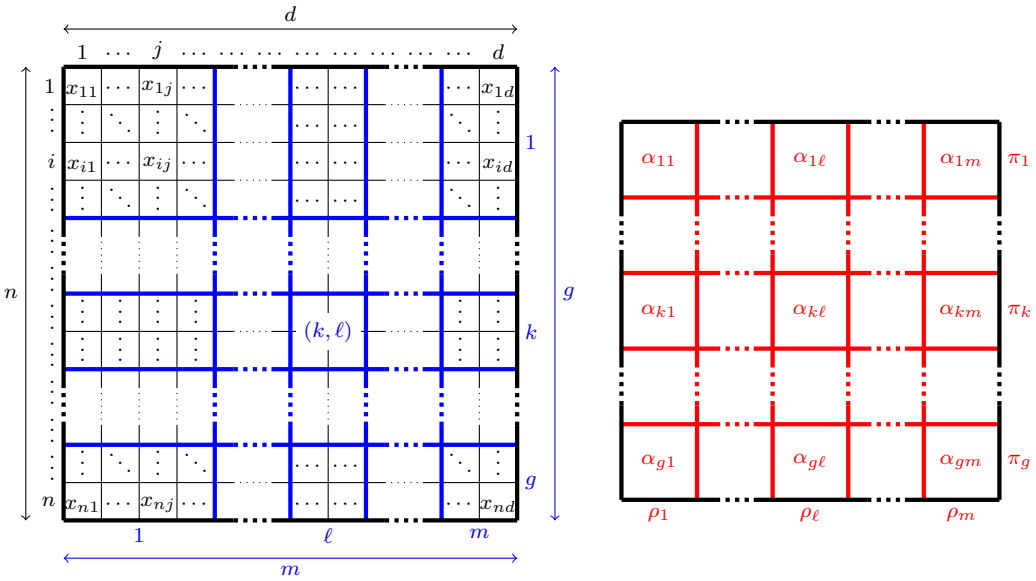


**Figure 1**. Notations. Left: Notations for the elements of observed data matrix are in black, notations for the block clusters are in blue. Right: Notations for the model parameter.

When performing inference from data, we note $\boldsymbol{\theta}^\star = (\boldsymbol{\pi}^\star, \boldsymbol{\rho}^\star, \boldsymbol{\alpha}^\star)$ 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) assignment of rows and columns to their group. For given matrices of indicator variables $\mathbf{z}$ and $\mathbf{w}$, we also note:

- $z_{+k} = \sum_i z_{ik}$ 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 to compare the partitions.

**Definition 2.1** (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* $\mathbf{w}^\star$*), noted* $\mathbb{R}_g(\mathbf{z})$ *(resp.* $\mathbb{R}_m(\mathbf{w})$*), as follows:*

$$\mathbb{R}_g(\mathbf{z})_{kk'} = \frac{1}{n}\sum_i z^\star_{ik} z_{ik'} \quad \textit{and} \quad \mathbb{R}_m(\mathbf{w})_{\ell\ell'} = \frac{1}{d}\sum_j w^\star_{j\ell} w_{j\ell'} \tag{2.1}$$

## 2.1. Likelihood

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

$$\begin{aligned}
\mathcal{L}_c(\mathbf{z}, \mathbf{w}; \boldsymbol{\theta}) &= \log p(\mathbf{x}, \mathbf{z}, \mathbf{w}; \boldsymbol{\theta}) \\
&= \log\left\{\left(\prod_{i,k} \pi_k^{z_{ik}}\right)\left(\prod_{j,\ell} \rho_\ell^{w_{j\ell}}\right)\left(\prod_{i,j,k,\ell} \varphi\left(x_{ij}; \alpha_{k\ell}\right)^{z_{ik}w_{j\ell}}\right)\right\} \\
&= \log\left\{\left(\prod_{i} \pi_{z_i}\right)\left(\prod_{j} \rho_{w_j}\right)\left(\prod_{i,j} \varphi\left(x_{ij}; \alpha_{z_i w_j}\right)\right)\right\}.
\end{aligned} \tag{2.2}$$

But the labels are usually unobserved, and the *observed log-likelihood* is obtained by marginalization over all the label configurations:

$$\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{x}; \boldsymbol{\theta}) = \log\left(\sum_{\mathbf{z}\in\mathcal{Z}, \mathbf{w}\in\mathcal{W}} p(\mathbf{x}, \mathbf{z}, \mathbf{w}; \boldsymbol{\theta})\right). \tag{2.3}$$

As the LBM involves a double missing data structure $\mathbf{z}$ for rows and $\mathbf{w}$ for columns, the observed likelihood is not tractable, nor the E-step of the EM algorithm, but estimation can be performed either by numerical approximation, or by MCMC methods [9], [8].

## 2.2. Assumptions

We focus here on parametric models where $\varphi$ belongs to a regular one-dimension exponential family in canonical form:

$$\varphi(x, \alpha) = b(x)\exp(\alpha x - \psi(\alpha)), \tag{2.4}$$

where $\alpha$ belongs to the space $\mathcal{A}$, so that $\varphi(\cdot, \alpha)$ is well defined for all $\alpha \in \mathcal{A}$. Classical properties of exponential families insure that $\psi$ is convex, infinitely differentiable on $\mathring{\mathcal{A}}$, that $(\psi')^{-1}$ is well defined on $\psi'(\mathring{\mathcal{A}})$. When $X_\alpha \sim \varphi(., \alpha)$, $\mathbb{E}[X_\alpha] = \psi'(\alpha)$ and $\mathbb{V}[X_\alpha] = \psi''(\alpha)$.

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

$H_1$ : There exist a positive constant $c$, and a compact $C_\alpha$ 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 map $\alpha \mapsto \varphi(\cdot, \alpha)$ is injective.
$H_4$ : Each row and each column of $\boldsymbol{\alpha}^\star$ is unique.

The previous assumptions are standard. Assumption $H_1$ ensure that the group proportions are bounded away from 0 and 1 so that no group disappears when $n$ and $d$ go to infinity. It also ensures that $\alpha$ is bounded away from the boundaries of the $\mathcal{A}$ and that there exists a $\kappa > 0$, such that $[\alpha_{k\ell} - \kappa, \alpha_{k\ell} + \kappa] \subset \mathring{\mathcal{A}}$ for all parameters $\alpha_{k\ell}$ of $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Assumptions $H_3$ and $H_4$ are necessary to ensure that the model is identifiable. If the map $\alpha \mapsto \varphi(., \alpha)$ is not injective, the model is trivially not identifiable. Similarly, if rows $k$ and $k'$ are identical, we can build a more parsimonious model that induces the same distribution of $\mathbf{x}$ by merging groups $k$ and $k'$. In the following, we consider that $g$ and $m$, row- and column- classes (or groups) counts are known.

Moreover, we define the $\delta(\boldsymbol{\alpha})$, that captures the differences between either row-groups or column-groups: lower values means that there are two row-classes or two column-classes that are very similar.

**Definition 2.2** (class distinctness)**.** *For $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha}) \in \boldsymbol{\Theta}$. We define:*

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

*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 $\varphi(., \alpha')$, when $\varphi$ comes from an exponential family.*

***Remark* 2.3.** Since all $\boldsymbol{\alpha}$ have distinct rows and columns, $\delta(\boldsymbol{\alpha}) > 0$.

***Remark* 2.4.** Since we restricted $\alpha$ in a bounded subset of $\mathring{\mathcal{A}}$, there exists two positive values $M_\alpha$ and $\kappa$ such that $C_\alpha + (-\kappa, \kappa) \subset [-M_\alpha, M_\alpha] \subset \mathring{\mathcal{A}}$. Moreover, the variance of $X_\alpha$ is bounded away from 0 and $+\infty$. We note

$$\sup_{\alpha \in [-M_\alpha, M_\alpha]} \mathbb{V}(X_\alpha) = \bar{\sigma}^2 < +\infty \quad \text{and} \quad \inf_{\alpha \in [-M_\alpha, M_\alpha]} \mathbb{V}(X_\alpha) = \underline{\sigma}^2 > 0. \tag{2.5}$$

**Proposition 2.5.** *With the previous notations, if $\alpha \in C_\alpha$ and $X_\alpha \sim \varphi(., \alpha)$, then $X_\alpha$ is subexponential with parameters $(\bar{\sigma}^2, \kappa^{-1})$.*

***Remark* 2.6.** 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 $\lambda$ 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 $\mu$, which is also the natural parameter, is bounded away from $\pm\infty$.

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

## 2.3. Symmetry

The LBM is a generalized mixture model, and it is well known that it subject to label switching. [9] showed that the categorical LBM is generically identifiable, and this property is easily extended to the case of observations of a one-dimension exponential family. Hence, except on a manifold set of null Lebesgue measure in $\boldsymbol{\Theta}$, the parameter set is identifiable up to a label permutation.

The study of the asymptotic properties of the MLE will lead to take into account symmetry properties on the parameter set. We first recall the definition of a permutation, then define equivalence relationships for assignments and parameter, and precise symmetry.

**Definition 2.7** (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 2.8** (equivalence)**.** *We define the following equivalence relationships:*

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

**Definition 2.9** (distance)**.** *We define the following distance, up to equivalence, between configurations* $\mathbf{z}$ *and* $\mathbf{z}^\star$*:*

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

*and similarly for the distance between* $\mathbf{w}$ *and* $\mathbf{w}^\star$ *where, for all matrix* $\mathbf{z}$*, we use the Hamming norm* $\|\cdot\|_0$ *defined by*

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

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 does not depend on $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$: 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 the confusion matrix $\mathbb{R}_g(\mathbf{z})$ (resp. $\mathbb{R}_m(\mathbf{w})$) is equivalent to a diagonal matrix.

**Definition 2.10** (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 noted* $\mathrm{Sym}(\boldsymbol{\theta})$*.*

***Remark* 2.11.** The set of parameters that exhibit symmetry is a manifold of null Lebesgue measure in $\boldsymbol{\Theta}$. The notion of symmetry allows us to deal with a notion of non-identifiability of the class labels that 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 $\boldsymbol{\alpha} = \begin{pmatrix} \alpha_1 & \alpha_2 \\ \alpha_2 & \alpha_1 \end{pmatrix}$ 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 exactly 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 $\theta$ exhibits symmetry, the maximum-likelihood *assignment* is not unique under the true model and there are at least $\#\,\mathrm{Sym}(\boldsymbol{\theta})$ of them.

## 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 (2.2):

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

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

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

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

$$\sqrt{nd}\left(\widehat{\alpha}_{k\ell}(\mathbf{z}^\star, \mathbf{w}^\star) - \alpha^\star_{k\ell}\right) \xrightarrow[n,d\to\infty]{\mathcal{D}} \mathcal{N}(0, \Sigma_{\boldsymbol{\alpha}^\star, k\ell}) \quad \textit{for all } k, \ell \tag{3.3}$$

*where the components are independent.*

*Proof:* Since $\hat{\boldsymbol{\pi}}(\mathbf{z}^\star) = (\hat{\pi}_1(\mathbf{z}^\star), \ldots, \hat{\pi}_g(\mathbf{z}^\star))$ (resp. $\hat{\boldsymbol{\rho}}(\mathbf{w}^\star)$) 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, $\psi'(\widehat{\alpha}_{k\ell}(\mathbf{z}^\star, \mathbf{w}^\star))$ is the average of $z^\star_{+k} w^\star_{+\ell} = nd\hat{\pi}_k(\mathbf{z}^\star)\widehat{\rho}_\ell(\mathbf{w}^\star)$ i.i.d. random variables with mean $\psi'(\alpha^\star_{k\ell})$ and variance $\psi''(\alpha^\star_{k\ell})$. $nd\hat{\pi}_k(\mathbf{z}^\star)\widehat{\rho}_\ell(\mathbf{w}^\star)$ is itself random but $\hat{\pi}_k(\mathbf{z}^\star)\widehat{\rho}_\ell(\mathbf{w}^\star) \xrightarrow[n,d\to+\infty]{} \pi^\star_k \rho^\star_\ell$ almost surely. Therefore, by Slutsky's lemma and the CLT for random sums of random variables [13], we have:

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

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

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

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

□

**Proposition 3.2** (Local asymptotic normality)**.** *Let $\mathcal{L}_c^\star$ the function defined on $\boldsymbol{\Theta}$ by $\mathcal{L}_c^\star(\boldsymbol{\pi},\boldsymbol{\rho},\boldsymbol{\alpha}) = \log p(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta})$. For any $s$, $t$ and $u$ in a compact set, we have:*

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

*where $\odot$ denote the Hadamard product of two matrices (element-wise product) and $\Sigma_{\boldsymbol{\pi}^\star}$, $\Sigma_{\boldsymbol{\rho}^\star}$ and $\Sigma_{\boldsymbol{\alpha}^\star}$ are defined in Proposition 3.1. $\mathbf{Y}_{\boldsymbol{\pi}^\star}$, $\mathbf{Y}_{\boldsymbol{\rho}^\star}$ are asymptotically Gaussian with zero mean and respective variance matrices $\Sigma_{\boldsymbol{\pi}^\star}$, $\Sigma_{\boldsymbol{\rho}^\star}$ and $\mathbf{Y}_{\boldsymbol{\alpha}^\star}$ is a matrix of asymptotically independent Gaussian components with zero mean and variance matrix $\Sigma_{\boldsymbol{\alpha}^\star}$.*

**Proof.**
By Taylor expansion,

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

where $\nabla\mathcal{L}_{c\,\boldsymbol{\pi}}^\star(\boldsymbol{\theta}^\star)$, $\nabla\mathcal{L}_{c\,\boldsymbol{\rho}}^\star(\boldsymbol{\theta}^\star)$ and $\nabla\mathcal{L}_{c\,\boldsymbol{\alpha}}^\star(\boldsymbol{\theta}^\star)$ 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 $\mathbf{H}_{\boldsymbol{\alpha}}$ denote 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 $\nabla\mathcal{L}_{c\,\boldsymbol{\pi}}^\star(\boldsymbol{\theta}^\star)/\sqrt{n}$, $\nabla\mathcal{L}_{c\,\boldsymbol{\rho}}^\star(\boldsymbol{\theta}^\star)/\sqrt{d}$ and $\nabla\mathcal{L}_{c\,\boldsymbol{\alpha}}^\star(\boldsymbol{\theta}^\star)/\sqrt{nd}$ converge in distribution by central limit theorem.

□

# 4. Profile Likelihood

To study the likelihood behaviors, we shall work conditionally to the real configurations $(\mathbf{z}^\star,\mathbf{w}^\star)$ that have enough observations in each row or column group. We therefore define regular configurations which occur with high probability, then introduce conditional and profile log-likelihood ratio.

## 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*

$$\min_k z_{+k} \geq cn \quad \text{and} \quad \min_\ell w_{+\ell} \geq cd. \tag{4.1}$$

In regular configurations, each row-group (resp. column-group) 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 $\boldsymbol{\theta}^\star \in \boldsymbol{\Theta}$.

Each $z_{+k}$ is a sum of $n$ i.i.d Bernoulli r.v. with parameter $\pi_k \geq \pi_{\min} \geq c$. A simple Hoeffding bound shows that

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

taking a union bound over $g$ values of $k$ and using a similar approach for $w_{+\ell}$ lead to 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 $H_1$. Note $\Omega_1$ the event $\{(\mathbf{z}^\star, \mathbf{w}^\star) \in \mathcal{Z}_1 \times \mathcal{W}_1\}$, then:*

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

We define now balls of configurations taking into account equivalent assignments classes.

**Definition 4.3** (Set of local assignments)**.** *We note $S(\mathbf{z}^\star, \mathbf{w}^\star, r)$ the set of configurations that have a representative (for $\sim$) within relative radius $r$ of $(\mathbf{z}^\star, \mathbf{w}^\star)$:*

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

## 4.2. Conditional and profile log-likelihoods

We first introduce few notations.

**Definition 4.4.** *We define the conditional log-likelihood ratio $F_{n,d}$ and its expectation $G$ as:*

$$\begin{aligned} F_{nd}(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}) &= \log\frac{p(\mathbf{x}|\mathbf{z}, \mathbf{w}; \boldsymbol{\theta})}{p(\mathbf{x}|\mathbf{z}^\star, \mathbf{w}^\star; \boldsymbol{\theta}^\star)} \\ G(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}) &= \mathbb{E}_{\boldsymbol{\theta}^\star}\left[\log\frac{p(\mathbf{x}|\mathbf{z}, \mathbf{w}; \boldsymbol{\theta})}{p(\mathbf{x}|\mathbf{z}^\star, \mathbf{w}^\star; \boldsymbol{\theta}^\star)}\middle| \mathbf{z}^\star, \mathbf{w}^\star\right] \end{aligned} \tag{4.2}$$

*We also define the profile log-likelihood ratio* $\Lambda$ *and its expectation* $\tilde{\Lambda}$ *as:*

$$\begin{aligned}\Lambda(\mathbf{z},\mathbf{w}) &= \max_{\boldsymbol{\theta}} F_{nd}(\boldsymbol{\theta},\mathbf{z},\mathbf{w})\\ \tilde{\Lambda}(\mathbf{z},\mathbf{w}) &= \max_{\boldsymbol{\theta}} G(\boldsymbol{\theta},\mathbf{z},\mathbf{w}).\end{aligned} \tag{4.3}$$

***Remark* 4.5.** 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 $\Lambda$ and $\tilde{\Lambda}$ allows us to get rid of the continuous argument of $F_{nd}$ and to effectively use discrete contrasts $\Lambda$ and $\tilde{\Lambda}$.

The following proposition shows which values of $\boldsymbol{\alpha}$ maximize $F_{nd}$ and $G$ to attain $\Lambda$ and $\tilde{\Lambda}$.

**Proposition 4.6** (maximum of $G$ and $\tilde{\Lambda}$ in $\theta$)**.** *Conditionally on* $\mathbf{z}^\star, \mathbf{w}^\star$*, define the following quantities:*

$$\begin{aligned}\boldsymbol{S}^\star &= (S^\star_{k\ell})_{k\ell} = (\psi'(\alpha^\star_{k\ell}))_{k\ell}\\ \bar{x}_{k\ell}(\mathbf{z},\mathbf{w}) &= \mathbb{E}_{\boldsymbol{\theta}^\star}[\widehat{x}_{k\ell}(\mathbf{z},\mathbf{w})|\mathbf{z}^\star,\mathbf{w}^\star] = \frac{\left[\mathbb{R}_g(\mathbf{z})^T \boldsymbol{S}^\star \mathbb{R}_m(\mathbf{w})\right]_{k\ell}}{\widehat{\pi}_k(\mathbf{z})\widehat{\rho}_\ell(\mathbf{w})}\end{aligned} \tag{4.4}$$

*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* $F_{nd}(\boldsymbol{\theta},\mathbf{z},\mathbf{w})$ *and* $G(\boldsymbol{\theta},\mathbf{z},\mathbf{w})$ *are maximum in* $\boldsymbol{\alpha}$ *for* $\widehat{\boldsymbol{\alpha}}(\mathbf{z},\mathbf{w})$ *and* $\bar{\boldsymbol{\alpha}}(\mathbf{z},\mathbf{w})$ *defined by:*

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

*so that*

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

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

The maximum and argmax of $G$ and $\tilde{\Lambda}$ are characterized by the following propositions.

**Proposition 4.7** (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:*

$$G(\boldsymbol{\theta},\mathbf{z},\mathbf{w}) = -nd\sum_{k,k'}\sum_{\ell,\ell'} \mathbb{R}_g(\mathbf{z})_{k,k'}\mathbb{R}_m(\mathbf{w})_{\ell,\ell'}\,\mathrm{KL}(\alpha^\star_{k\ell},\alpha_{k'\ell'}) \leq 0. \tag{4.5}$$

*Conditionally on the set* $\Omega_1$ *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; moreover,* $\tilde{\Lambda}(\mathbf{z}^\star, \mathbf{w}^\star) = 0$.
*(iii) The maximum of* $\tilde{\Lambda}$ *(and hence the maximum of* $G$*) is well separated.*

Property *(iii)* of Proposition 4.7 is a direct consequence of the local upperbound for $\tilde{\Lambda}$ as stated as follows:

**Proposition 4.8** (Local upperbound for $\tilde{\Lambda}$)**.** *Conditionally upon* $\Omega_1$*, there exists a positive constant* $C$ *such that for all* $(\mathbf{z}, \mathbf{w}) \in S(\mathbf{z}^\star, \mathbf{w}^\star, C)$*:*

$$\tilde{\Lambda}(\mathbf{z}, \mathbf{w}) \leq -\frac{c\delta(\boldsymbol{\alpha}^\star)}{4} \left(d\|\mathbf{z} - \mathbf{z}^\star\|_{0,\sim} + n\|\mathbf{w} - \mathbf{w}^\star\|_{0,\sim}\right) \tag{4.6}$$

Proofs of Propositions 4.6, 4.7 and 4.8 are reported in Appendix A.

# 5. Main Result

We are now ready to present our main result stated in Theorem 5.1.

**Theorem 5.1** (complete-observed)**.** *Consider that assumptions* $H_1$ *to* $H_4$ *hold for the Latent Block Model of known order with* $n \times d$ *observations coming from an univariate exponential family and define* $\#\,\mathrm{Sym}(\boldsymbol{\theta})$ *as the set of pairs of permutation* $(s, t)$ *for which* $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha})$ *exhibits symmetry. Then, for* $n$ *and* $d$ *tending to infinity with asymptotic rates* $\log(d)/n \to 0$ *and* $\log(n)/d \to 0$*, the observed likelihood ratio behaves like the complete likelihood ratio, up to a bounded multiplicative factor:*

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

*where the* $o_P$ *is 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 already pointed out, if $\boldsymbol{\Theta}$ exhibits symmetry, the maximum likelihood assignment is not unique under the true model, and $\#\,\mathrm{Sym}(\boldsymbol{\theta})$ terms contribute with the same weight. This was not taken into account by [2]. The next corollary is deduced immediately :

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

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

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

Using the conditional log-likelihood, the observed likelihood can be written as

$$\begin{aligned} p(\mathbf{x};\boldsymbol{\theta}) &= \sum_{(\mathbf{z},\mathbf{w})} p(\mathbf{x},\mathbf{z},\mathbf{w};\boldsymbol{\theta}) \\ &= p(\mathbf{x}|\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}^\star) \sum_{(\mathbf{z},\mathbf{w})} p(\mathbf{z},\mathbf{w};\boldsymbol{\theta}) \exp(F_{nd}(\boldsymbol{\theta},\mathbf{z},\mathbf{w})). \end{aligned} \tag{5.1}$$

The proof proceeds with an examination of the asymptotic behavior of $F_{nd}$ on three types of configurations that partition $\mathcal{Z}\times\mathcal{W}$:

1. *global control*: for $(\mathbf{z},\mathbf{w})$ such that $\tilde{\Lambda}(\mathbf{z},\mathbf{w}) = \Omega(-nd)$, Proposition 5.3 proves a large deviation behavior for $F_{nd} = -\Omega_P(nd)$ and in turn those assignments contribute a $o_P$ of $p(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}^\star))$ to the sum (Proposition 5.4).
2. *local control*: a small deviation result (Proposition 5.5) 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 $o_P$ of $p(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}^\star)$ (Proposition 5.6).
3. *equivalent assignments*: Proposition 5.7 examines which of the remaining assignments, all equivalent to $(\mathbf{z}^\star,\mathbf{w}^\star)$, contribute to the sum.

These results are presented in next section 5.1 and their proofs reported in Appendix A. They are then put together in section 5.2 to achieve the proof of our main result. The remainder of the section is devoted to the asymptotics of the ML and variational estimators as a consequence of the main result.

## 5.1. Different asymptotic behaviors

We begin with a large deviations inequality for configurations $(\mathbf{z},\mathbf{w})$ far from $(\mathbf{z}^\star,\mathbf{w}^\star)$ and leverage it to prove that far away configurations make a small contribution to $p(\mathbf{x};\boldsymbol{\theta})$.

### *5.1.1. Global Control*

**Proposition 5.3** (large deviations of $F_{nd}$)**.** *Let* $\mathrm{Diam}(\boldsymbol{\Theta}) = \sup_{\boldsymbol{\theta},\boldsymbol{\theta}'} \|\boldsymbol{\theta}-\boldsymbol{\theta}'\|_\infty$. *For all* $\varepsilon_{n,d} < \kappa/(2\sqrt{2}\,\mathrm{Diam}(\boldsymbol{\Theta}))$ *and* $n, d$ *large enough that*

$$\begin{aligned} &\Delta^1_{nd}(\varepsilon_{nd}) \\ &\quad = \mathbb{P}\left(\sup_{\boldsymbol{\theta},\mathbf{z},\mathbf{w}} \left\{F_{nd}(\boldsymbol{\theta},\mathbf{z},\mathbf{w}) - \tilde{\Lambda}(\mathbf{z},\mathbf{w})\right\} \geq \bar{\sigma} nd\,\mathrm{Diam}(\boldsymbol{\Theta}) 2\sqrt{2}\varepsilon_{nd}\left[1+\frac{gm}{2\sqrt{2nd}\varepsilon_{nd}}\right]\right) \\ &\qquad\qquad \leq g^n m^d \exp\left(-\frac{nd\varepsilon^2_{nd}}{2}\right) \end{aligned} \tag{5.2}$$

**Proposition 5.4** (contribution of global assignments)**.** *Assume* $\log(d)/n \to 0$, $\log(n)/d \to 0$ *when* $n$ *and* $d$ *tend to infinity, and choose* $t_{nd}$ *decreasing to* 0 *such that* $t_{nd} \gg \max(\frac{n+d}{nd}, \frac{\log(nd)}{\sqrt{nd}})$.

*Then conditionally on $\Omega_1$ and for $n, d$ large enough that $2\sqrt{2ndt}_{nd} \geq gm$, we have:*

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

*5.1.2. Local Control*

Proposition 5.3 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 5.5** (small deviations $F_{nd}$)**.** *Conditionally upon $\Omega_1$, there exists three positive constant $c_1$, $c_2$ and $C$ such that for all $\varepsilon \leq \kappa\underline{\sigma}^2$, for all $(\mathbf{z},\mathbf{w}) \nsim (\mathbf{z}^\star,\mathbf{w}^\star)$ such that $(\mathbf{z},\mathbf{w}) \in S(\mathbf{z}^\star,\mathbf{w}^\star,C)$:*

$$\Delta_{nd}^2(\varepsilon) = \mathbb{P}_{\boldsymbol{\theta}^\star}\left(\sup_{\boldsymbol{\theta}} \frac{F_{nd}(\boldsymbol{\theta},\mathbf{z},\mathbf{w}) - \tilde{\Lambda}(\mathbf{z},\mathbf{w})}{d\|\mathbf{z}-\mathbf{z}^\star\|_{0,\sim} + n\|\mathbf{w}-\mathbf{w}^\star\|_{0,\sim}} \geq \varepsilon\right) \leq \exp\left(-\frac{ndc^2\varepsilon^2}{128(c_1\bar{\sigma}^2 + c_2\kappa^{-1}\varepsilon)}\right) \tag{5.3}$$

The next propositions builds on Proposition 5.5 and 4.7 to show that the combined contributions of assignments close to $(\mathbf{z}^\star,\mathbf{w}^\star)$ to the observed likelihood is also a $o_P$ of $p(\mathbf{z}^\star,\mathbf{w}^\star,\mathbf{x};\boldsymbol{\theta}^\star)$

**Proposition 5.6** (contribution of local assignments)**.** *With the previous notations*

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

*5.1.3. Equivalent assignments*

It remains to study the contribution of equivalent assignments.

**Proposition 5.7** (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 $\boldsymbol{\theta}$.*

## 5.2. Proof of the main result

**Proof.**
We work conditionally on $\Omega_1$. 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 $t_{nd} \gg \max\left(\frac{n+d}{nd}, \frac{\log(nd)}{\sqrt{nd}}\right)$ (this is possible since $\log(d)/n \to 0$ and $\log(n)/d \to 0$). According to Proposition 5.4,

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

Since $t_{nd}$ decreases to 0, it gets smaller than $C$ (used in proposition 5.6) for $n, d$ large enough. As this point, Proposition 5.6 ensures that:

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

And therefore the observed likelihood ratio reduces as:

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

And Proposition 5.7 allows us to conclude

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

□

## 5.3. Asymptotics for the MLE of $\boldsymbol{\theta}$

The asymptotic behavior of the maximum likelihood estimator in the incomplete data model is a direct consequence of Theorem 5.1.

**Corollary 5.8** (Asymptotic behavior of $\widehat{\boldsymbol{\theta}}_{MLE}$)**.** *Denote $\widehat{\boldsymbol{\theta}}_{MLE}$ the maximum likelihood estimator and use the notations of Proposition 3.1. If $\#\,\mathrm{Sym}(\boldsymbol{\theta}) = 1$, there exist permutations $s$ of $\{1,\dots,g\}$ and $t$ of $\{1,\dots,m\}$ such that*

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

*If* $\#\,\mathrm{Sym}(\boldsymbol{\theta}) \neq 1$, $\widehat{\boldsymbol{\theta}}_{MLE}$ *is still consistent: there exist permutations* $s$ *of* $\{1,\dots,g\}$ *and* $t$ *of* $\{1,\dots,m\}$ *such that*

$$\begin{aligned} \hat{\boldsymbol{\pi}}(\mathbf{z}^\star) - \widehat{\boldsymbol{\pi}}^s_{MLE} = o_P(1), \qquad & \hat{\boldsymbol{\rho}}(\mathbf{w}^\star) - \widehat{\boldsymbol{\rho}}^t_{MLE} = o_P(1), \\ \hat{\boldsymbol{\alpha}}(\mathbf{z}^\star, \mathbf{w}^\star) - \widehat{\boldsymbol{\alpha}}^{s,t}_{MLE} &= o_P(1). \end{aligned}$$

Hence, the maximum likelihood estimator for the LBM is consistent and asymptotically normal, with the same behavior as the maximum likelihood estimator in the complete data model when $\theta$ does not exhibit any symmetry. The proof in appendix A.9 relies on the local asymptotic normality of the MLE in the complete model, as stated in Proposition 3.2 and on our main Theorem.

## 5.4. Consistency of variational estimates

Due to the complex dependence structure of the observations, the maximum likelihood estimator of the LBM is not numerically tractable, even with the *Expectation Maximisation* algorithm. In practice, a variational approximation can be used ([**?**, see for example]]govaert2003): 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 product distributions, such that for all $(\mathbf{z}, \mathbf{w})$

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

allow to obtain tractable expressions of $J(\mathbb{Q}, \boldsymbol{\theta})$. The variational estimate $\widehat{\boldsymbol{\theta}}_{var}$ of $\boldsymbol{\theta}$ is defined as

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

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

**Corollary 5.9** (Variational estimate)**.** *Under the assumptions of Theorem 5.1 and if* $\#\,\mathrm{Sym}(\boldsymbol{\theta}) = 1$*, there exist permutations* $s$ *of* $\{1,\dots,g\}$ *and* $t$ *of* $\{1,\dots,m\}$ *such that*

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

The proof is available in appendix A.10.

# Appendix A: Proofs

## A.1. Proof of Proposition 4.6

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

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

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

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

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

□

## A.2. Proof of Proposition 4.7 (maximum of $G$ and $\tilde{\Lambda}$)

**Proof.**
We condition on $(\mathbf{z}^\star, \mathbf{w}^\star)$ and prove Equation (4.5):

$$\begin{aligned}
G(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}) &= \mathbb{E}_{\boldsymbol{\theta}^\star} \left[ \frac{p(\mathbf{x}; \mathbf{z}, \mathbf{w}, \boldsymbol{\theta})}{p(\mathbf{x}; \mathbf{z}^\star, \mathbf{w}^\star, \boldsymbol{\theta}^\star)} \middle| \mathbf{z}^\star, \mathbf{w}^\star \right] \\
&= \sum_i \sum_j \sum_{k,k'} \sum_{\ell,\ell'} \mathbb{E}_{\boldsymbol{\theta}^\star} \left[ x_{ij}(\alpha_{k'\ell'} - \alpha^\star_{k\ell}) - (\psi(\alpha_{k'\ell'}) - \psi(\alpha^\star_{k\ell})) \right] z^\star_{ik} z_{ik'} w^\star_{j\ell} w_{j\ell'} \\
&= nd \sum_{k,k'} \sum_{\ell,\ell'} \mathbb{R}_g(\mathbf{z})_{k,k'} \mathbb{R}_m(\mathbf{w})_{\ell,\ell'} \left[ \psi'(\alpha^\star_{k\ell})(\alpha_{k'\ell'} - \alpha^\star_{k\ell}) + \psi(\alpha^\star_{k\ell}) - \psi(\alpha_{k'\ell'}) \right] \\
&= -nd \sum_{k,k'} \sum_{\ell,\ell'} \mathbb{R}_g(\mathbf{z})_{k,k'} \mathbb{R}_m(\mathbf{w})_{\ell,\ell'} \, \mathrm{KL}(\alpha^\star_{k\ell}, \alpha_{k'\ell'})
\end{aligned}$$

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 B.4 (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 4.8.

□

## A.3. Proof of Proposition 4.8 (Local upper bound for $\tilde{\Lambda}$)

**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 computing the gradient of $\tilde{\Lambda}$ at and around its argmax and using those gradients to control the local behavior of $\tilde{\Lambda}$ around its argmax. The local behavior allows us 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 $\mathbb{R}_g(\mathbf{z})\mathbb{1} = \boldsymbol{\pi}(\mathbf{z}^\star)$ and $\mathbb{R}_m(\mathbf{w})\mathbb{1} = \boldsymbol{\rho}(\mathbf{w}^\star)$, 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 $(D_\pi, D_\rho) \coloneqq (\mathrm{Diag}(\boldsymbol{\pi}(\mathbf{z}^\star)), \mathrm{Diag}(\boldsymbol{\rho}(\mathbf{w}^\star)))$ is:

$$A_{kk'}(\mathbf{w}^\star) \coloneqq \frac{\partial \tilde{\Lambda}}{\partial U_{kk'}}(D_\pi, D_\rho) = \sum_{\ell} \rho_\ell(\mathbf{w}^\star) \, \mathrm{KL}\left(\alpha^\star_{k\ell}, \alpha^\star_{k'\ell}\right)$$

$$B_{\ell\ell'}(\mathbf{z}^\star) \coloneqq \frac{\partial \tilde{\Lambda}}{\partial V_{\ell\ell'}}(D_\pi, D_\rho) = \sum_{k} \rho_\ell(\mathbf{z}^\star) \, \mathrm{KL}\left(\alpha^\star_{k\ell}, \alpha^\star_{k\ell'}\right)$$

$A(\mathbf{w}^\star)$ and $B(\mathbf{z}^\star)$ are the matrix-derivative of $-\tilde{\Lambda}/nd$ at $(D_\pi, D_\rho)$. 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. $\ell$). By boundedness of the second derivative, there exists $C > 0$ such that for all $(D_\pi, D_\rho)$ and all $(H,G) \in B(D_\pi, D_\rho, C)$, we have:

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

Choose $U$ and $V$ in $(\mathcal{U}_c \times \mathcal{V}_c) \cap B(D_\pi, D_\rho, C)$ 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 $U, V, D_\pi, D_\rho$ 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 $(\mathcal{U}_c \times \mathcal{V}_c) \cap B(D_\pi, D_\rho, C)$ such that

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

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 $(\mathcal{U}_c \times \mathcal{V}_c) \cap B(D_\pi, D_\rho, C)$ 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.

□

The maps $f_{k,\ell} : x \mapsto KL(\alpha^\star_{k\ell}, (\psi')^{-1}(x))$ are twice differentiable with a continuous second derivative bounded by $\underline{\sigma}^{-2}$ on $\psi'(C_\alpha)$. All terms $\left[U^T \boldsymbol{S}^\star V\right]_{k\ell}\left[U^T \mathbf{1} V\right]^{-1}_{k\ell}$ are convex combinations of the $\psi'(\alpha^\star_{k\ell})$ and therefore in $\psi'(C_\alpha)$. Furthermore, their first and second order derivative are also bounded as soon as each row sum of $U$ and $V$ is bounded away from 0. By composition, all second order partial derivatives of $\tilde{\Lambda}$ are therefore continuous and bounded on $\mathcal{U} \times \mathcal{V}$.

We now compute the first derivative of $\tilde{\Lambda}$ at $(D_\pi, D_\rho) \coloneqq (\mathrm{Diag}(\boldsymbol{\pi}(\mathbf{z}^\star)), \mathrm{Diag}(\boldsymbol{\rho}(\mathbf{w}^\star)))$ by doing a first-order Taylor expansion of $\tilde{\Lambda}\left(D_\pi + U, D_\rho + V\right)$ for small $U$ and $V$.

Tedious but straightforward manipulations show:

$$\begin{aligned}\bar{\alpha}_{k\ell}(D_\pi + U, D_\rho + V) =& \alpha^\star_{k\ell} + \frac{1}{\pi_k(\mathbf{z}^\star)}\sum_{k'} U_{kk'}(S_{k'\ell} - 1)\\ &+ \frac{1}{\rho_\ell(\mathbf{w}^\star)}\sum_{\ell'} V_{\ell\ell'}(S_{k\ell'} - 1) + o(\|U\|_1, \|V\|_1)\\ \mathrm{KL}\left(\alpha^\star_{k\ell}, \bar{\alpha}_{k'\ell'}\right) =& \mathrm{KL}\left(\alpha^\star_{k\ell}, \alpha^\star_{k',\ell'}\right) + \begin{cases}\mathcal{O}(\|U\|_1, \|V\|_1) & \text{if } (k',\ell') \neq (k,\ell)\\ o(\|U\|_1, \|V\|_1) & \text{if } (k',\ell') = (k,\ell)\end{cases}\end{aligned}$$

where the second line comes from the fact that $f'_{k,\ell}(\psi'(\alpha^\star_{k\ell})) = 0$. Keeping only the first order term in $U$ and $V$ in $\tilde{\Lambda}$ and noting that $\tilde{\Lambda}\left(D_\pi, D_\rho\right) = 0$ yields:

$$\frac{-1}{nd}[\tilde{\Lambda}\left(D_\pi + U, D_\rho + V\right) - \tilde{\Lambda}\left(D_\pi, D_\rho\right)] = \frac{-1}{nd}\tilde{\Lambda}\left(D_\pi + U, D_\rho + V\right)$$
$$= \sum_k D_{\pi,kk} \sum_{\ell,\ell'} V_{\ell\ell'} \,\mathrm{KL}\left(\alpha^\star_{k\ell}, \bar{\alpha}_{k\ell'}\right) + \sum_\ell D_{\rho,\ell\ell} \sum_{k,k'} U_{kk'} \,\mathrm{KL}\left(\alpha^\star_{k\ell}, \bar{\alpha}_{k'\ell}\right) + o(\|U\|_1, \|V\|_1)$$
$$= \sum_k \pi_k(\mathbf{z}^\star) \sum_{\ell,\ell'} V_{\ell\ell'} \,\mathrm{KL}\left(\alpha^\star_{k\ell}, \alpha^\star_{k\ell'}\right) + \sum_\ell \rho_\ell(\mathbf{w}^\star) \sum_{k,k'} U_{kk'} \,\mathrm{KL}\left(\alpha^\star_{k\ell}, \alpha^\star_{k'\ell}\right) + o(\|U\|_1, \|V\|_1)$$
$$= \mathrm{Tr}(UA(\mathbf{w}^\star)) + \mathrm{Tr}(VB(\mathbf{z}^\star)) + o(\|U\|_1, \|V\|_1)$$

where $A_{kk'}(\mathbf{w}^\star) \coloneqq \sum_\ell \rho_\ell(\mathbf{w}^\star) \,\mathrm{KL}\left(\alpha^\star_{k\ell}, \alpha^\star_{k'\ell}\right)$ and $B_{\ell\ell'}(\mathbf{z}^\star) \coloneqq \sum_k \rho_\ell(\mathbf{z}^\star) \,\mathrm{KL}\left(\alpha^\star_{k\ell}, \alpha^\star_{k\ell'}\right)$. $A$ and $B$ are the matrix-derivative of $-\tilde{\Lambda}/nd$ at $(D_\pi, D_\rho)$. Since $(\mathbf{z}^\star, \mathbf{w}^\star)$ is $c/2$-regular and by definition of $\delta(\boldsymbol{\alpha}^\star)$, $A_{kk'} \geq c\delta(\boldsymbol{\alpha}^\star)/2$ for $k \neq k'$ and $B_{\ell\ell'} \geq c\delta(\boldsymbol{\alpha}^\star)/2$ for $\ell \neq \ell'$ and the diagonal terms of $A$ and $B$ are null. By boundedness of the lower second derivative of $\tilde{\Lambda}$, there exists a constant $C > 0$ such that for all $(H, G) \in B(D_\pi, D_\rho, C)$, we have:

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

In particular, if $U$ and $V$ have nonnegative non diagonal coefficients and negative diagonal coefficients.

$$\frac{-1}{nd}\left[\mathrm{Tr}\left(U\frac{\partial\tilde{\Lambda}}{\partial U}(H,G)\right) + \mathrm{Tr}\left(V\frac{\partial\tilde{\Lambda}}{\partial V}(H,G)\right)\right]$$
$$\geq \frac{c\delta(\boldsymbol{\alpha}^\star)}{4}\left[\sum_{kk'} U_{kk'} + \sum_{\ell\ell'} V_{\ell\ell'} - \mathrm{Tr}(U) - \mathrm{Tr}(V)\right]$$

Choose $U$ and $V$ in $(\mathcal{U}\times\mathcal{V}) \cap B(D_\pi, D_\rho, c_3)$ satisfying $U\mathbb{1} = \boldsymbol{\pi}(\mathbf{z}^\star)$ and $V\mathbb{1} = \boldsymbol{\rho}(\mathbf{w}^\star)$. Note that $U - D_\pi$ and $V - D_\rho$ have nonnegative non diagonal coefficients, negative diagonal coefficients, that their coefficients sum up to 1 and that $\mathrm{Tr}(D_\pi) = \mathrm{Tr}(D_\rho) = 1$. By Taylor expansion, there exists a couple $(H, G)$ also in $(\mathcal{U} \times \mathcal{V}) \cap B(D_\pi, D_\rho, C)$ such that

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

To conclude the proof, choose any assignment $(\mathbf{z}, \mathbf{w})$ and without loss of generality assume that $(\mathbf{z}, \mathbf{w})$ are closest to $(\mathbf{z}^\star, \mathbf{w}^\star)$ in their equivalence class. Then $\mathbb{R}_g(\mathbf{z})$ is in $\mathcal{U}$ and additionally satifies $\mathbb{R}_g(\mathbf{z})\mathbb{1} = \boldsymbol{\pi}(\mathbf{z}^\star)$ and $\|\mathbf{z} - \mathbf{z}^\star\|_{0,\sim} = n\|\mathbb{R}_g(\mathbf{z}) - D_\pi\|_1/2 = n(1 - \mathrm{Tr}(\mathbb{R}_g(\mathbf{z})))$. Similar equalities hold for $\mathbb{R}_m(\mathbf{w})$ and $\|\mathbf{w} - \mathbf{w}^\star\|_0$.

□

## A.4. Proof of Proposition 5.3 (global convergence $F_{nd}$)

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

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

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 B.3, $\mathbb{E}_{\boldsymbol{\theta}^\star}[Z|\mathbf{z}^\star, \mathbf{w}^\star] \leq gm\,\mathrm{Diam}(\boldsymbol{\Theta})\sqrt{nd\bar{\sigma}^2}$ 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 $\varepsilon_{n,d} < \kappa/2\sqrt{2}\,\mathrm{Diam}(\boldsymbol{\Theta})$

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

We can then remove the conditioning and take a union bound to prove Equation (5.2).

□

## A.5. Proof of Proposition 5.4 (contribution of far away assignments)

**Proof.**
Conditionally on $(\mathbf{z}^\star, \mathbf{w}^\star)$, we know from proposition 4.7 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 $t_{nd} \gg \max\left(\frac{n+d}{nd}, \frac{\log(nd)}{\sqrt{nd}}\right)$. This is always possible because we assume that $\log(d)/n \to 0$ and $\log(n)/d \to 0$. According to 4.7 (iii), for all $(\mathbf{z}, \mathbf{w}) \notin (\mathbf{z}^\star, \mathbf{w}^\star, t_{nd})$

$$\tilde{\Lambda}(\mathbf{z}, \mathbf{w}) \leq -\frac{c\delta(\boldsymbol{\alpha}^\star)}{4}(n\|\mathbf{w} - \mathbf{w}^\star\|_{0,\sim} + d\|\mathbf{z} - \mathbf{z}^\star\|_{0,\sim}) \leq -\frac{c\delta(\boldsymbol{\alpha}^\star)}{4}ndt_{nd} \tag{A.1}$$

since either $\|\mathbf{z} - \mathbf{z}^\star\|_{0,\sim} \geq nt_{nd}$ or $\|\mathbf{w} - \mathbf{w}^\star\|_{0,\sim} \geq dt_{nd}$.

Set $\varepsilon_{nd} = \frac{\inf(c\delta(\boldsymbol{\alpha}^\star)t_{nd}/16\bar{\sigma}, \kappa)}{\mathrm{Diam}(\boldsymbol{\Theta})}$. By proposition 5.3, and with our choice of $\varepsilon_{nd}$, with probability higher than $1 - \Delta^1_{nd}(\varepsilon_{nd})$,

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

where the second line comes from inequality (A.1), the third from the global control studied in Proposition 5.3 and the definition of $\varepsilon_{nd}$, 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 $\varepsilon_{nd} \gg \log(nd)/\sqrt{nd}$ so that the series $\sum_{n,d} \Delta^1_{nd}(\varepsilon_{nd})$ converges and:

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

□

## A.6. Proof of Proposition 5.5 (local convergence $F_{nd}$)

**Proof.**
We work conditionally on $(\mathbf{z}^\star, \mathbf{w}^\star) \in \mathcal{Z}_1 \times \mathcal{W}_1$. Choose $\varepsilon \leq \kappa\underline{\sigma}^2$ small. Assignments $(\mathbf{z}, \mathbf{w})$ at $\|.\|_{0,\sim}$-distance less than $c/4$ of $(\mathbf{z}^\star, \mathbf{w}^\star)$ are $c/4$-regular. According to Proposition B.1, $\widehat{x}_{k\ell}$ and $\bar{x}_{k\ell}$ are at distance at most $\varepsilon$ with probability higher than $1 - \exp\left(-\frac{ndc^2\varepsilon^2}{128(\bar{\sigma}^2+\kappa^{-1}\varepsilon)}\right)$. Manipulation of $\Lambda$ and $\tilde{\Lambda}$ yield

$$\begin{aligned}\frac{F_{nd}(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}) - \tilde{\Lambda}(\mathbf{z}, \mathbf{w})}{nd} &\leq \frac{\Lambda(\mathbf{z}, \mathbf{w}) - \tilde{\Lambda}(\mathbf{z}, \mathbf{w})}{nd} \\ &= \sum_{k,k''} \sum_{\ell,\ell'} \mathbb{R}_g(\mathbf{z})_{kk'} \mathbb{R}_m(\mathbf{w})_{\ell\ell'} \left[f_{k\ell}(\widehat{x}_{k'\ell'}) - f_{k\ell}(\bar{x}_{k'\ell'})\right]\end{aligned}$$

where $f_{k\ell}(x) = -S^\star_{k\ell}(\psi')^{-1}(x) + \psi \circ (\psi')^{-1}(x)$. The functions $f_{k\ell}$ are twice differentiable on $\mathring{\mathcal{A}}$ with bounded first and second derivatives over $I = \psi'([-M_\alpha, M_\alpha])$ so that:

$$f_{k\ell}(y) - f_{k\ell}(x) = f'_{k\ell}(x)\,(y - x) + o\,(y - x)$$

where the $o$ is uniform over pairs $(x, y) \in I^2$ at distance less than $\varepsilon$ and does not depend on $(\mathbf{z}^\star, \mathbf{w}^\star)$. $\bar{x}_{k\ell}$ is a convex combination of the $S^\star_{k\ell} = \psi'(\alpha^\star_{k\ell}) \in \psi'(C_\alpha)$. Since $\psi'$ is monotonic, $\bar{x}_{k\ell} \in \psi'(C_\alpha) \subset I$. Similarly, $|\widehat{x}_{k\ell} - \bar{x}_{k\ell}| \leq \kappa\underline{\sigma}^2$ and $|\psi''| \geq \underline{\sigma}^2$ over $I$ therefore $\widehat{x}_{k\ell} \in I$. We now bound $f'_{k\ell}$:

$$\begin{aligned}|f'_{k\ell}(\bar{x}_{k'\ell'})| = \left|\frac{\bar{x}_{k'\ell'} - S^\star_{k\ell}}{\psi'' \circ (\psi')^{-1}(\bar{x}_{k'\ell'})}\right| &= \left|\frac{\frac{\left[\mathbb{R}_g(\mathbf{z})^T \boldsymbol{S}^\star \mathbb{R}_m(\mathbf{w})\right]_{k'\ell'}}{\widehat{\pi}_k(\mathbf{z})\widehat{\rho}_\ell(\mathbf{w})} - S^\star_{k\ell}}{\psi'' \circ (\psi')^{-1}(\bar{x}_{k'\ell'})}\right| \\ &\leq \left(1 - \frac{\mathbb{R}_g(\mathbf{z})_{kk'}\mathbb{R}_m(\mathbf{w})_{\ell\ell'}}{\widehat{\pi}_k(\mathbf{z})\widehat{\rho}_\ell(\mathbf{w})}\right) \frac{S^\star_{\max} - S^\star_{\min}}{\underline{\sigma}^2}\end{aligned}$$

where $S^\star_{\max} = \max_{k,\ell} S^\star_{k\ell}$ and $S^\star_{\min} = \min_{k,\ell} S^\star_{k\ell}$. In particular,

$$\begin{aligned}\mathbb{R}_g(\mathbf{z})_{kk'}\mathbb{R}_m(\mathbf{w})_{\ell\ell'}|f'_{k\ell}(\bar{x}_{k'\ell'})| &\leq \mathbb{R}_g(\mathbf{z})_{kk'}\mathbb{R}_m(\mathbf{w})_{\ell\ell'}\left(1 - \frac{\mathbb{R}_g(\mathbf{z})_{kk'}\mathbb{R}_m(\mathbf{w})_{\ell\ell'}}{\widehat{\pi}_k(\mathbf{z})\widehat{\rho}_\ell(\mathbf{w})}\right)\frac{S^\star_{\max} - S^\star_{\min}}{\underline{\sigma}^2} \\ &\leq \begin{cases}\mathbb{R}_g(\mathbf{z})_{kk'}\mathbb{R}_m(\mathbf{w})_{\ell\ell'}\frac{S^\star_{\max}-S^\star_{\min}}{\underline{\sigma}^2} & \text{if } (k',\ell') \neq (k,\ell) \\ [\widehat{\pi}_k(\mathbf{z})\widehat{\rho}_\ell(\mathbf{w}) - \mathbb{R}_g(\mathbf{z})_{kk}\mathbb{R}_m(\mathbf{w})_{\ell\ell}]\frac{S^\star_{\max}-S^\star_{\min}}{\underline{\sigma}^2} & \text{if } (k,\ell) = (k,\ell)\end{cases}\end{aligned}$$

Wrapping everything,

$$
\begin{aligned}
\frac{|\Lambda(\mathbf{z},\mathbf{w})-\tilde{\Lambda}(\mathbf{z},\mathbf{w})|}{nd} &= \left|\sum_{k,k'}\sum_{\ell,\ell'} \mathbb{R}_g(\mathbf{z})_{kk'}\mathbb{R}_m(\mathbf{w})_{\ell\ell'}\left[f'_{k\ell}(\bar{x}_{k'\ell'})(\widehat{x}_{k\ell}-\bar{x}_{k\ell})+o(\widehat{x}_{k\ell}-\bar{x}_{k\ell})\right]\right| \\
&\leq \left[\sum_{(k',\ell')\neq(k,\ell)} \mathbb{R}_g(\mathbf{z})_{kk'}\mathbb{R}_m(\mathbf{w})_{\ell\ell'} + \sum_{k,\ell}(\widehat{\pi}_k(\mathbf{z})\widehat{\rho}_\ell(\mathbf{w}) - \mathbb{R}_g(\mathbf{z})_{kk}\mathbb{R}_m(\mathbf{w})_{\ell\ell}\right] \\
&\times \frac{S^\star_{\max}-S^\star_{\min}}{\underline{\sigma}^2}\max_{k,\ell}|\widehat{x}_{k\ell}-\bar{x}_{k\ell}|(1+o(1)) \\
&= 2\left[\sum_{(k',\ell')\neq(k,\ell)} \mathbb{R}_g(\mathbf{z})_{kk'}\mathbb{R}_m(\mathbf{w})_{\ell\ell'}\right]\frac{S^\star_{\max}-S^\star_{\min}}{\underline{\sigma}^2}\max_{k,\ell}|\widehat{x}_{k\ell}-\bar{x}_{k\ell}|(1+o(1)) \\
&= 2\left[1-\operatorname{Tr}(\mathbb{R}_g(\mathbf{z}))\operatorname{Tr}(\mathbb{R}_m(\mathbf{w}))\right]\frac{S^\star_{\max}-S^\star_{\min}}{\underline{\sigma}^2}\max_{k,\ell}|\widehat{x}_{k\ell}-\bar{x}_{k\ell}|(1+o(1)) \\
&\leq 2\left(\frac{\|\mathbf{z}-\mathbf{z}^\star\|}{n}+\frac{\|\mathbf{w}-\mathbf{w}^\star\|}{d}\right)\frac{S^\star_{\max}-S^\star_{\min}}{\underline{\sigma}^2}\max_{k,\ell}|\widehat{x}_{k\ell}-\bar{x}_{k\ell}|(1+o(1)) \\
&\leq 2\left(\frac{\|\mathbf{z}-\mathbf{z}^\star\|}{n}+\frac{\|\mathbf{w}-\mathbf{w}^\star\|}{d}\right)\frac{S^\star_{\max}-S^\star_{\min}}{\underline{\sigma}^2}\varepsilon(1+o(1))
\end{aligned}
$$

We can remove the conditioning on $(\mathbf{z}^\star,\mathbf{w}^\star)$ to prove Equation (5.3) with $c_2 = 2(S^\star_{\max}-S^\star_{\min})/\underline{\sigma}^2$ and $c_1 = c_2^2$.

□

## A.7. Proof of Proposition 5.6 (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 $\Omega_1$. We work conditionally on $(\mathbf{z}^\star,\mathbf{w}^\star)\in\mathcal{Z}_1\times\mathcal{W}_1$. Choose $(\mathbf{z},\mathbf{w})$ in $S(\mathbf{z}^\star,\mathbf{w}^\star,C)$ with $C$ defined in proposition 5.4.

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

For $C$ small enough, we can assume without loss of generality that $(\mathbf{z},\mathbf{w})$ is the representative closest to $(\mathbf{z}^\star,\mathbf{w}^\star)$ and note $r_1 = \|\mathbf{z}-\mathbf{z}^\star\|_0$ and $r_2 = \|\mathbf{w}-\mathbf{w}^\star\|_0$. We choose

$\varepsilon_{nd} \leq \min(\kappa\underline{\sigma}^2, c\delta(\boldsymbol{\alpha}^\star)/8)$. Then with probability at least $1 - \exp\left(-\frac{nd\tilde{c}^2\varepsilon_{nd}^2}{8(c_1\bar{\sigma}^2 + c_2\kappa^{-1}\varepsilon_{nd})}\right)$:

$$\begin{aligned} F_{nd}(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}) &\leq \Lambda(\mathbf{z}, \mathbf{w}) - \tilde{\Lambda}(\mathbf{z}, \mathbf{w}) + \tilde{\Lambda}(\mathbf{z}, \mathbf{w}) \\ &\leq \Lambda(\mathbf{z}, \mathbf{w}) - \tilde{\Lambda}(\mathbf{z}, \mathbf{w}) - \frac{c\delta(\boldsymbol{\alpha}^\star)}{4}(dr_1 + nr_2) \\ &\leq \varepsilon_{nd}(dr_1 + nr_2) - \frac{c\delta(\boldsymbol{\alpha}^\star)}{4}(dr_1 + nr_2) \\ &\leq -\frac{c\delta(\boldsymbol{\alpha}^\star)}{8}(dr_1 + nr_2) \end{aligned}$$

where the first line comes from the definition of $\Lambda$, the second line from Proposition 4.7, the third from Proposition 5.5 and the last from $\varepsilon_{nd} \leq c\delta(\boldsymbol{\alpha}^\star)/8$. A union bound shows that

$$\begin{aligned} \Delta_{nd}(\varepsilon_{nd}) = \mathbb{P}_{\boldsymbol{\theta}^\star}\left(\sup_{\substack{(\mathbf{z},\mathbf{w})\in(\mathbf{z}^\star,\mathbf{w}^\star,\tilde{c}) \\ \boldsymbol{\theta}\in\boldsymbol{\Theta}}} F_{nd}(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}) \geq -\frac{c\delta(\boldsymbol{\alpha}^\star)}{8}(d\|\mathbf{z} - \mathbf{z}^\star\|_{0,\sim} + n\|\mathbf{w} - \mathbf{w}^\star\|_{0,\sim})\right) \\ \leq g^n m^d \exp\left(-\frac{nd\tilde{c}^2\varepsilon_{nd}^2}{8(c_1\bar{\sigma}^2 + c_2\kappa^{-1}\varepsilon_{nd})}\right) \end{aligned}$$

Thanks to corollary B.6, we also know that:

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

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, with probability $1 - \Delta_{nd}(\varepsilon_{nd})$,

$$\begin{aligned} &\frac{\sum_{\substack{(\mathbf{z},\mathbf{w})\in S(\mathbf{z}^\star,\mathbf{w}^\star,\tilde{c}) \\ (\mathbf{z},\mathbf{w})\nsim(\mathbf{z}^\star,\mathbf{w}^\star)}} p(\mathbf{z}, \mathbf{w}, \mathbf{x}; \boldsymbol{\theta})}{p(\mathbf{z}^\star, \mathbf{w}^\star, \mathbf{x}; \boldsymbol{\theta}^\star)} \\ &\leq \mathcal{O}_P(1)\sum_{r_1+r_2\geq 1}\binom{n}{r_1}\binom{n}{r_2}g^{g+r_1}m^{m+r_2}\exp\left((r_1 + r_2)M_{c/4} - \frac{c\delta(\boldsymbol{\alpha}^\star)}{8}(dr_1 + nr_2)\right) \\ &= \mathcal{O}_P(1)\left(1 + e^{(g+1)\log g + M_{c/4} - d\frac{c\delta(\boldsymbol{\alpha}^\star)}{8}}\right)^n\left(1 + e^{(m+1)\log m + M_{c/4} - n\frac{c\delta(\boldsymbol{\alpha}^\star)}{8}}\right)^d - 1 \\ &\leq \mathcal{O}_P(1)a_{nd}\exp(a_{nd}) \end{aligned}$$

where $a_{nd} = ne^{(g+1)\log g + M_{c/4} - d\frac{c\delta(\boldsymbol{\alpha}^\star)}{8}} + de^{(m+1)\log m + M_{c/4} - n\frac{c\delta(\boldsymbol{\alpha}^\star)}{8}} = o(1)$ as soon as $n \gg \log d$ and $d \gg \log n$. If we take $\varepsilon_{nd} \gg \log(nd)/\sqrt{nd}$, the series $\sum_{n,d}\Delta_{nd}(\varepsilon_{nd})$ converges which proves the results.

□

## A.8. Proof of Proposition 5.7 (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

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

$p(\mathbf{x},\mathbf{z}^{\star},\mathbf{w}^{\star};\boldsymbol{\theta})$ unimodal in $\boldsymbol{\theta}$, with a mode in $\widehat{\boldsymbol{\theta}}_{MC}$. By consistency of $\widehat{\boldsymbol{\theta}}_{MC}$, 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 $p(\mathbf{x},\mathbf{z}^{\star},\mathbf{w}^{\star};\boldsymbol{\theta}) = \mathcal{O}_P(p(\mathbf{x},\mathbf{z}^{\star},\mathbf{w}^{\star};\boldsymbol{\theta}^{\star}))$ and $\boldsymbol{\theta} \to \boldsymbol{\theta}^{\star}$. 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,

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

□

## A.9. Proof of Corollary 5.8: Behavior of $\widehat{\boldsymbol{\theta}}_{MLE}$

Theorem 5.1, states that:

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

Then,

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

Now, using Corollary 3 p. 553 of Mariadassou and Matias [11]

$$p(\cdot,\cdot|\mathbf{x};\boldsymbol{\theta}^\star) \xrightarrow[n,d\to+\infty]{(\mathcal{D})} \frac{1}{\#\,\mathrm{Sym}(\boldsymbol{\theta}^\star)} \sum_{(\mathbf{z},\mathbf{w})\overset{\boldsymbol{\theta}^\star}{\sim}(\mathbf{z}^\star,\mathbf{w}^\star)} \delta_{(\mathbf{z},\mathbf{w})}(\cdot,\cdot),$$

we can deduce that

$$\begin{aligned} p(\mathbf{x};\boldsymbol{\theta}) &= \#\,\mathrm{Sym}(\boldsymbol{\theta})\frac{1}{\#\,\mathrm{Sym}(\boldsymbol{\theta}^\star)p(\mathbf{z}^\star,\mathbf{w}^\star|\mathbf{x};\boldsymbol{\theta}^\star)} \max_{\boldsymbol{\theta}'\sim\boldsymbol{\theta}} p(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}')\,(1+o_P(1))+o_P(1) \\ &= \#\,\mathrm{Sym}(\boldsymbol{\theta})\frac{1}{1+o_P(1)} \max_{\boldsymbol{\theta}'\sim\boldsymbol{\theta}} p(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}')\,(1+o_P(1))+o_P(1) \\ &= \#\,\mathrm{Sym}(\boldsymbol{\theta}) \max_{\boldsymbol{\theta}'\sim\boldsymbol{\theta}} p(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}')\,(1+o_P(1))+o_P(1). \end{aligned} \tag{A.2}$$

Finaly, we conclude with the proposition 3.2.

## A.10. Proof of Corollary 5.9: Behavior of $J(\mathbb{Q},\boldsymbol{\theta})$

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

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

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

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

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

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

Moreover, using Equation A.2, we get a lower bound:

$$\begin{aligned} \max_{(\mathbf{z},\mathbf{w})\in\mathcal{Z}\times\mathcal{W}} \frac{p(\mathbf{x},\mathbf{z},\mathbf{w};\boldsymbol{\theta})}{p(\mathbf{x};\boldsymbol{\theta}^\star)} &= \max_{\boldsymbol{\theta}'\sim\boldsymbol{\theta}} \frac{p\left(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}'\right)(1+o_p(1))}{p(\mathbf{x};\boldsymbol{\theta}^\star)} + o_p(1) \\ &= \max_{\boldsymbol{\theta}'\sim\boldsymbol{\theta}} \frac{p\left(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}'\right)(1+o_p(1))}{\#\,\mathrm{Sym}(\boldsymbol{\theta}^\star)p(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}^\star)(1+o_p(1))} + o_p(1) \\ &= \max_{\boldsymbol{\theta}'\sim\boldsymbol{\theta}} \frac{p\left(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}'\right)(1+o_p(1))}{\#\,\mathrm{Sym}(\boldsymbol{\theta}^\star)p(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}^\star)} + o_p(1). \end{aligned}$$

Now, Theorem 5.1 leads to the following upper bound:

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

so that we have the following control

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

In the particular case where $\#\,\mathrm{Sym}(\boldsymbol{\theta})=1$, we have

$$\frac{\max\limits_{\mathbb{Q}\in\mathcal{Q}} \exp\left[J\left(\mathbb{Q},\boldsymbol{\theta}\right)\right]}{p\left(\mathbf{x};\boldsymbol{\theta}^\star\right)} = \frac{1}{\#\,\mathrm{Sym}(\boldsymbol{\theta}^\star)}\max_{\boldsymbol{\theta}'\sim\boldsymbol{\theta}}\frac{p\left(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}'\right)\left(1+o_p(1)\right)}{p\left(\mathbf{x},\mathbf{z}^\star,\mathbf{w}^\star;\boldsymbol{\theta}^\star\right)}+o_p(1)$$

and, following the same reasoning as the appendix A.9, we have the result.

# Appendix B: Technical Lemma

## B.1. Sub-exponential variables

We now prove two propositions regarding subexponential variables. Recall first that a random variable $X$ is sub-exponential with parameters $(\tau^2, b)$ if for all $\lambda$ 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:

$$\mathbb{P}(X-\mathbb{E}[X]\geq t) \leq \begin{cases}\exp\left(-\frac{t^2}{2\tau^2}\right) & \text{if} \quad 0\leq t\leq \frac{\tau^2}{b} \\ \exp\left(-\frac{t}{2b}\right) & \text{if} \quad t\geq \frac{\tau^2}{b}\end{cases} \tag{B.1}$$

So that

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

The subexponential property is preserved by summation and multiplication.

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

**Proposition B.1** (Maximum in $(\mathbf{z}, \mathbf{w})$)**.** *Let $(\mathbf{z}, \mathbf{w})$ be a configuration and $\hat{x}_{k,\ell}(\mathbf{z}, \mathbf{w})$ resp. $\bar{x}_{k\ell}(\mathbf{z}, \mathbf{w})$ be as defined in Equations (3.1) and (4.4). Under the assumptions of the section 2.2, for all $\varepsilon > 0$*

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

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

$$\mathbb{P}\left(\max_{\mathbf{z}\in\mathcal{Z}_1,\mathbf{w}\in\mathcal{W}_1} \max_{k,l} |\hat{x}_{k,\ell} - \bar{x}_{k\ell}| > \varepsilon\right) \leq g^{n+1}m^{d+1} \exp\left(-\frac{ndc^2\varepsilon^2}{8(\bar{\sigma}^2 + \kappa^{-1}\varepsilon)}\right). \quad \text{(B.3)}$$

Note that equations B.2 and B.3 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 $(\mathbf{z}^\star, \mathbf{w}^\star)$, $z_{+k}w_{+\ell}(\hat{x}_{k,\ell} - \bar{x}_{k\ell})$ is a sum of $z_{+k}w_{+\ell}$ centered subexponential random variables. By Bernstein's inequality [12], we therefore have for all $t > 0$

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

In particular, if $t = ndx$,

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

uniformly over $(\mathbf{z}, \mathbf{w})$. Equation (B.2) then results from a union bound. Similarly,

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

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 (B.3) then results from a union bound over $\mathcal{Z}_1 \times \mathcal{W}_1 \subset \mathcal{Z} \times \mathcal{W}$.

□

**Lemma B.2.** *If $X$ is a zero mean random variable, subexponential with parameters $(\sigma^2, b)$, then $|X|$ is subexponential with parameters $(8\sigma^2, 2\sqrt{2}b)$.*

**Proof.**
Note $\mu = \mathbb{E}|X|$ and consider $Y = |X| - \mu$. Choose $\lambda$ 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 subexponential 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

$$\begin{aligned}
\mathbb{E}[e^{\lambda Y}] &\leq 1 + \left(\frac{1}{2} + \frac{1}{2.3!}\right)\mathbb{E}[Y^2]\lambda^2 + \sum_{k=2}^{+\infty}\left(\frac{1}{(2k)!} + \frac{1}{2}\left[\frac{1}{(2k-1)!} + \frac{1}{(2k+1)!}\right]\right)\lambda^{2k}\mathbb{E}[Y^{2k}] \\
&\leq \sum_{k=0}^{+\infty} 2^k \frac{\lambda^{2k}\mathbb{E}[Y^{2k}]}{(2k)!} \\
&\leq \sum_{k=0}^{+\infty} 2^{3k} \frac{\lambda^{2k}\mathbb{E}[X^{2k}]}{(2k)!} = \cosh\left(2\sqrt{2}\lambda X\right) = \mathbb{E}\left[\frac{e^{2\sqrt{2}\lambda X} + e^{-2\sqrt{2}\lambda X}}{2}\right] \\
&\leq e^{\frac{8\lambda^2\sigma^2}{2}}
\end{aligned}$$

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 $\mathbb{E}[Y^{2k}]$.

□

**Proposition B.3** (concentration for subexponential)**.** *Let $X_1, \ldots, X_n$ be independent zero mean random variables, subexponential with parameters $(\sigma_i^2, b_i)$. Note $V_0^2 = \sum_i \sigma_i^2$ 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 subexponential with parameters* $(8M^2V_0^2, 2\sqrt{2}Mb)$. *Moreover* $\mathbb{E}[Z] \leq MV_0\sqrt{n}$ *so that for all* $t > 0$,

$$\mathbb{P}(Z - MV_0\sqrt{n} \geq t) \leq \exp\left(-\frac{t^2}{2(8M^2V_0^2 + 2\sqrt{2}Mbt)}\right) \tag{B.4}$$

**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 B.2 and standard properties of sums of independent rescaled subexponential variables.

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

using Cauchy-Schwarz.

□

The final lemma is the working horse for proving Proposition 4.7.

**Lemma B.4.**
*Let* $\eta$ *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* $\eta$ *are distinct;*
- *all the columns* $\eta$ *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 note*

$$\Sigma = \sum_{kk'}\sum_{\ell\ell'} A_{kk'}B_{\ell\ell'}d_{k\ell k'\ell'} \tag{B.5}$$

*Then,*

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

**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. $\ell$), the following sum $\Sigma$ reduces to

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

$\Sigma$ 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 $(k, \ell)$, $d_{s(k)t(\ell)k\ell} = 0$ 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 $A_{k_1 k}$ and $A_{k_2 k}$. Let $\ell \in \{1 \dots m\}$, there exists by assumption $\ell'$ such that $B_{\ell\ell'} \neq 0$. As $\Sigma = 0$, both products $A_{k_1 k} B_{\ell\ell'} d_{k_1 \ell k\ell'}$ and $A_{k_2 k} B_{\ell\ell'} d_{k_2 \ell k\ell'}$ are zero.

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

The previous equality is true for all $\ell$, thus rows $k_1$ and $k_2$ of $\eta$ are identical, and contradict the assumptions.

□

## B.2. Likelihood ratio of assignments

**Lemma B.5.**
*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 note $\mathbb{S}^g_c = \mathbb{S}^g \cap [c, 1-c]^g$. 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 $\boldsymbol{\pi} \in \mathbb{S}^g_c$*

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

**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 $\nabla H$ 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 $\hat{\boldsymbol{\pi}}(\mathbf{z}) \in \mathbb{S}^g \cap [c, 1-c]^g$, that $\log p(\mathbf{z}; \hat{\boldsymbol{\pi}}(\mathbf{z})) - \log p(\mathbf{z}^\star; \hat{\boldsymbol{\pi}}(\mathbf{z}^\star)) = n[H(\hat{\boldsymbol{\pi}}(\mathbf{z})) - H(\hat{\boldsymbol{\pi}}(\mathbf{z}^\star))]$ and finally that $\|\hat{\boldsymbol{\pi}}(\mathbf{z}) - \hat{\boldsymbol{\pi}}(\mathbf{z}^\star)\|_1 \leq \frac{2}{n} \|\mathbf{z} - \mathbf{z}^\star\|_0$.

□

**Corollary B.6.** *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:

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

where the first inequality comes from the definition of $\hat{\boldsymbol{\pi}}(\mathbf{z})$ and $\hat{\boldsymbol{\rho}}(\mathbf{w})$ and the second from Lemma B.5 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 $\frac{p(\mathbf{z}^\star, \mathbf{w}^\star; \hat{\boldsymbol{\pi}}(\mathbf{z}^\star), \hat{\boldsymbol{\rho}}(\mathbf{w}^\star))}{p(\mathbf{z}^\star, \mathbf{w}^\star; \boldsymbol{\pi}^\star, \boldsymbol{\rho}^\star)} = \mathcal{O}_P(1)$.

□

# 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.

[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.

[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.

[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.

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

[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.

[10] Aurore Lomet. *Sélection de modèles pour la classification de données continues*. PhD thesis, Université Technologique de Compiègne, 2012.

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

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

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