Regression with Label Permutation in Generalized Linear Model

Guanhua Fang
School of Management, Fudan University
670 Guoshun Road, Shanghai 200433, China
fanggh@fudan.edu.cn

Ping Li
LinkedIn Ads
700 Bellevue Way NE, Bellevue, WA 98004, USA
pinli@linkedin.com

Abstract

The assumption that response and predictor belong to the same statistical unit may be violated in practice. Unbiased estimation and recovery of true label ordering based on unlabeled data are challenging tasks and have attracted increasing attentions in the recent literature. In this paper, we present a relatively complete analysis of label permutation problem for the generalized linear model with multivariate responses. The theory is established under different scenarios, with knowledge of true parameters, with partial knowledge of underlying label permutation matrix and without any knowledge. Our results remove the stringent conditions required by the current literature and are further extended to the missing observation setting which has never been considered in the field of label permutation problem. On computational side, we propose two methods, "maximum likelihood estimation" algorithm and "two-step estimation" algorithm, to accommodate for different settings. When the proportion of permuted labels is moderate, both methods work effectively. Multiple numerical experiments are provided and corroborate our theoretical findings.

1. Introduction

A key assumption in regression problems is that response-predictor pairs correspond to the same statistical unit. In practice, this assumption may be violated when different subsets of variables are collected asynchronously and are merged together with certain label disagreements. That is, responses and predictors may not be perfectly paired together so that the statistical inferences based on such label-contaminated data sets could be inaccurate and biased. Research on the unlabeled problem has a long history and can be traced back to 1970s under the name "broken sample problem" (DeGroot et al., 1971; Goel, 1975; DeGroot and Goel, 1976; 1980; Chan and Loh, 2001; Bai and Hsing, 2005; Slawski et al., 2020). In recent years, we have witnessed a renaissance of this problem due to its wide applications, such as data integration, privacy protection, computer vision, robotics, sensor networks, etc. See Unnikrishnan et al. (2015); Pananjady et al. (2018; 2017); Slawski and Ben-David (2019); Zhang et al. (2019); Slawski et al. (2020); Zhang et al. (2022); Zhang and Li (2023a) and the references therein for more explanations.

Important applications of label permutation include linkage record, data de-anonymization, and header-free communication. In linkage record (Newcombe and Kennedy, 1962; Fellegi and Sunter, 1969), people would like to integrate multiple databases, where each contains different pieces of information about the same identity, into one comprehensive database. In this process, the biggest challenge is how to find the matching across different databases. For data de-anonymization (Nazarov et al., 2018), the task is to identify the labels, which aims to preserve privacy, with public data. It can be seen as the inverse problem of privacy protection. For the header-free communication (Pananjady et al., 2017; Shi and Shi, 2019), we have a sensor network where the sensor identity is omitted during communication to reduce the transmission cost and latency. In this scenario, reconstruction of signal involves recovering the unknown correspondence.

In the literature, for the sake of simplicity, linear models are often considered for studying the label permutation problem. However, the linear model assumption may be violated when the error distribution is skewed or heavy tailed. Additionally, linear models cannot capture the structure of count data which is another popular data type and is becoming increasingly ubiquitous (e.g. survival analysis (Fleming and Harrington, 2011), online streaming services (Cugola and Margara, 2012), educational testing (Templin and Henson, 2010), etc). With these in mind, in this paper, we specifically adopt the formulation of generalized linear model (GLM) to take care of multivariate

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

non-Gaussian responses. The problem is formulated as,

$$Y = \Pi^\natural Y^\natural, \tag{1}$$

where $Y^\natural$ is the response matrix when the data are labeled correctly. For each entry of $Y^\natural$, $Y^\natural[i,l]$ admits the density

$$f_{il}(y) = \exp\{y\lambda_{il} - \psi(\lambda_{il}) + c(y)\}$$

for $i \in [n]$ and $l \in [m]$ with $\lambda_{il} = \mathbf{x}_i^T \mathbf{b}_l^\natural$. Here, n is the number of units/individuals, m is the number of observations for each unit/individual, λ_{il} refers to the natural parameter. $c(y)$ is a nuisance function free of parameter. (E.g. $c(y) = \exp\{-y^2/2\}/\sqrt{2\pi}$ for standard normal density; $c(y) = \log(y!)$ for Poisson density.) Unobserved $\Pi^\natural \in \mathbb{R}^{n\times n}$ denotes an underlying row permutation matrix. In other words, we only observe the response matrix up to certain label permutations. $X := (\mathbf{x}_i) \in \mathbb{R}^{n\times p}$ represents the covariate/design matrix which is fully observed; $B^\natural := (\mathbf{b}_l^\natural) \in \mathbb{R}^{p\times m}$ is the underlying true parameter coefficient matrix, which may be unknown. Our task is to recover the label permutation matrix $\Pi^\natural$ based on permuted data Y and design matrix X.

Related work. There is a rapidly growing body of literature on regression problems with unknown label permutation, starting from Unnikrishnan et al. (2015; 2018); Pananjady et al. (2018). Paper (Pananjady et al., 2018) presents necessary and sufficient conditions for permutation recovery for linear models with Gaussian design. Extensions to multivariate linear models are considered in Pananjady et al. (2017); Zhang et al. (2019); Zhang and Li (2020); Zhang et al. (2022); Zhang and Li (2023a). The papers (Abid et al., 2017; Hsu et al., 2017) show that consistent estimation of the regression parameter is impossible without substantial additional assumptions. Tsakiris et al. (2018); Tsakiris (2018) have studied important theoretical aspects such as well-posedness from an algebraic perspective, and have also put forth practical computational schemes such as a branch-and-bound algorithm (Emiya et al., 2014) and concave maximization (Peng and Tsakiris, 2020). An approximate EM scheme with a Markov-Chain-Monte-Carlo (MCMC) approximation of the E-step is discussed in Abid and Zou (2018). Additionally, approaches to linear and multivariate linear regression with sparsely mismatched data are studied in Slawski and Ben-David (2019); Slawski et al. (2020; 2021; 2019). A tight analyses on sparse regression problem is provided in Zhang and Li (2021; 2023b). Nevertheless, on the other hand, a relatively small amount of papers have considered regression with unlabeled/permuted data outside the standard linear model. The topics of those papers include spherical regression (Shi et al., 2020), univariate isotonic regression and statistical seriation (Carpentier and Schlueter, 2016; Rigollet and Weed, 2019; Flammarion et al., 2019; Ma et al., 2020; Balabdaoui et al., 2021), and binary regression (Wang et al., 2018).

The most related work is Wang et al. (2020), where they consider a generalized linear regression models within exponential family. The difference is that they only consider the case $m = 1$ and the corresponding theoretical analyses are established when the true parameter $B^\natural$ is assumed to be known. The case $m > 1$ should be of independent interest for the following reasons. First, in the context of record linkage, it is natural to assume that both predictor matrix X and response Y are multi-dimensional. Second, the availability of multiple responses affected by the same permutation is expected to facilitate estimation. Stronger assumptions are required for label recovery when $m = 1$, while such assumptions can be relaxed when m grows as n grows. In addition, the estimation problem is more challenging when the true parameter $B^\natural$ is unknown. The theory under unknown parameter settings is waiting to be developed. To be reader-friendly, Table 1 summarizes the key differences of our setting from the existing ones.

	Linear model	GLM
Uni-dim ($m = 1$)	Pananjady et al. (2018) Unnikrishnan et al. (2018)	Wang et al. (2020)
Multi-dim ($m > 1$)	Zhang et al. (2019) Zhang and Li (2020) Slawski et al. (2020)	Ours

Table 1. A categorization summary of literature review.

Contributions. In this paper, we study the label permutation problem under generalized linear model framework which is different from the classical linear model in the following ways. The response could be discrete instead of continuous such that the resulting estimator does not admit a nice closed form. In the linear model, the signal-to-noise ratio (SNR) plays an important role in recovering the underlying labels. In contrast, there is no such unified criteria in the generalized linear model. The existing work of Wang et al. 2020 for generalized linear model only considers the case of $m = 1$. When the regression parameter B is assumed to be known, the sufficient condition for label permutation recovery requires $\min_{1\le i_1\ne i_2\le n} |\psi'(\lambda_{i_1}) - \psi'(\lambda_{i_2})|$ grows as n grows. There are no theoretical guarantees under the situation $m > 1$ or the situation when regression parameter B is unknown. In this paper, we bridge this gap and show the perfect label recovery results under different scenarios. Moreover, we also consider the situation of missing observations, i.e., each entry in response Y may be missing completely at random with certain probability. The corresponding theory has also been established in this paper. Such missing observation case has not been studied yet in the literature for unlabeled regression problems. On the technical side, we also want to point out the analysis of generalized linear models is harder than that of linear models due to the existence of exponential function whose second order derivative may not be bounded. Additionally, we need to take special care of analyzing the maximum likelihood estimator which admits no closed form and involves $n!$ different possible permutations. The results in current paper add theoretical values in many applications, e.g., data integration, privacy protection, etc.

Outline. The rest of paper is organized as follows. Section 2 introduces the generalized linear model setting and useful notations. Section 3 presents the permutation recovery analysis when parameters are assumed to be known. In Section 4, the main theory is established when parameters are unknown and underlying permutation matrix is partially known or unknown. Moreover, we further extend the results to the missing observation setting in Section 5. The concluding remarks are given in Section 6. Numerical results and technical proofs are deferred to appendices.

2. Permutation Problem

2.1. Toy example

An illustrative example is shown in Table 2. On the left side, it presents the personal information (i.e. salary and age) of five individuals in the correct label order. On the right side, information of salary and age is coupled in a wrong label order due to the privacy reasons or errors caused by merging data from different sources. Obviously, salary and age are **impossible** to follow Gaussian distributions. Our goal is to recover the true label order given the permuted data set and under certain model assumptions in the framework of generalized linear model.

Table 2. A toy example of label permutation problem.

Original			Permuted	
Label	Salary	Age	Salary (Label)	Age (Label)
1	6500	50	6500 (1)	45 (3)
2	4300	30	5000 (3)	30 (2)
3	5000	45	3200 (4)	50 (1)
4	3200	25	4300 (2)	25 (4)
5	8000	55	8000 (5)	55 (5)

2.2. Model

In this paper, we specifically study the generalized linear model with label permutation. The problem can be written in the following matrix form,

$$Y = \Pi^{\natural} Y^{\natural}, \tag{2}$$

where $Y^{\natural} \sim f(XB^{\natural})$, i.e., $Y^{\natural}[i,l] \sim f_{il}(y)$ with

$$f_{il}(y) = \exp\{y\lambda_{il} - \psi(\lambda_{il}) + c(y)\},\ \lambda_{il} = \mathbf{x}_i^T \mathbf{b}_l^{\natural}$$

for $i \in [n]$ and $l \in [m]$. Unobserved $\Pi^{\natural} \in \mathbb{R}^{n\times n}$ denotes an underlying row permutation matrix (i.e. a binary matrix with each row/column containing one and only one non-zero entry), $X \in \mathbb{R}^{n\times p}$ represents the covariate/design matrix, and $B^{\natural} \in \mathbb{R}^{p\times m}$ is the underlying true parameter coefficient matrix. Here $\psi(\cdot)$ is a smooth univariate convex function over $\mathbb{R}$. When we take function $\psi(x) = x^2$, then the density is a normal density. When we take $\psi(x) = \exp\{x\}$, then it becomes a Poisson distribution. Our goal is to recover the underlying permutation matrix $\Pi^{\natural}$ given mislabeled observations Y and X.

For a fixed permutation Π, the log-likelihood function after removing the nuisance parts (the term not related to the parameters) is given by

$$\begin{aligned} L(\Pi, B) &= \sum_{i=1}^{n}\sum_{l=1}^{m}\left\{Y[i,l](\mathbf{x}_{\Pi(i)}^T\mathbf{b}_l) - \psi(\mathbf{x}_{\Pi(i)}^T\mathbf{b}_l)\right\} \\ &= \langle -\psi(\Pi XB) + Y \circ \Pi XB\rangle. \end{aligned} \tag{3}$$

In the rest of paper, we consider to recover $\Pi^{\natural}$ by maximizing the above log-likelihood function. When the true parameter matrix $B^{\natural}$ is known, the estimator will be $\hat{\Pi} := \arg\max_{\Pi} L(\Pi, B^{\natural})$. When the parameter matrix B is unknown, the estimator will be

$$\hat{\Pi} := \arg\max_{\Pi}\max_{B} L(\Pi, B).$$

Notation. We use $\natural$ to denote the true value and use $a \gtrsim b$ ($a \lesssim b$) to represent $a \geq Kb$ ($a \leq b/K$) for some sufficiently large constant K. $\|\mathbf{a}\|$ and $\|A\|$ represent ℓ_2 norm of vector $\mathbf{a}$ and spectral norm of matrix A, respectively. For any $\mathcal{S}$, its cardinality is denoted by $|\mathcal{S}|$. For two positive real numbers a and b, $b = O(a), b = \Omega(a)$ and $b = \Theta(a)$ indicate the relations, $b \leq C_2 a$, $b \geq a/C_1$ and $a/C_1 \leq b \leq C_2 a$, correspondingly, where C_1, C_2 are some constants. For random sequences, $x_n = o_p(1)$ means x_n converges to 0 in probability and $x_n = O_p(y_n)$ means x_n/y_n is stochastically bounded as $n \to \infty$. For an arbitrary univariate function f, $f(\mathbf{a})$ and $f(A)$ are obtained via applying f to vector $\mathbf{a}$ and matrix A elementwisely. We let $\psi'(\lambda)$ and $\psi''(\lambda)$ be the first and second order derivative of $\psi(\lambda)$. Generic constants c, c_0, c_1, C, c_{ψ} may vary from place to place. A complete notation list is given in Table 3.

3. Recovery Analysis when $B^{\natural}$ is known

3.1. Permutation Recovery

When $B^{\natural}$ is known, we only need to estimate Π by maximizing the likelihood function without estimating B. In other words, the best estimator should be

$$\begin{aligned} \hat{\Pi} &:= \arg\max_{\Pi} L(\Pi, B^{\natural}) \\ &= \arg\max_{\Pi}\langle -\psi(\Pi XB^{\natural}) + Y \circ \Pi XB^{\natural}\rangle. \end{aligned}$$

Table 3. Notation List.

Notation	Definition
n	the number of individuals
m	the number of observations for each individual
p	the number of covariates/predictors
X	the covariate/design matrix (n by p)
Y	the observed/response matrix (n by m)
$\mathbf{y}_i$	the response vector for individual i
B	the coefficient/parameter matrix
$\mathbf{x}_i^T$	the ith row of X
$\mathbf{b}_l$	the lth column of B
Π	the row permutation matrix (n by n)
$\mathbf{I}$	the identity of row permutation matrix
$\Pi(i)$	the permuted label for individual i
$d(\Pi_1, \Pi_2)$	the Hamming distance, i.e., $\sum_{i=1}^n \mathbf{1}\{\Pi_1(i) \neq \Pi_2(i)\}$
$\boldsymbol{\lambda}_i$	$(\mathbf{x}_i^T\mathbf{b}_1, \ldots, \mathbf{x}_i^T\mathbf{b}_m)$, i.e., the vector of linear component for individual i
$\langle \mathbf{a} \rangle$ / $\langle A \rangle$	the sum of all entries in vector $\mathbf{a}$ / matrix A
$\mathbf{a}[l]$	the lth element of vector $\mathbf{a}$
$A[i,j]$	the element of matrix A in ith row and jth column
$A[\mathcal{S},:]/A[:,\mathcal{S}]$	the sub-matrix of A with row/column indices from set $\mathcal{S}$
$A_1 \circ A_2$	the Hadamard product of A_1 and A_2
$[K]$	$\{1, \ldots, K\}$ for any positive integer K.
$\\|\mathbf{x}\\|$ / $\\|\mathbf{x}\\|_1$	ℓ_2-norm / ℓ_1-norm for vector $\mathbf{x}$.
$\|\Omega\|$	the cardinality of set Ω.

Successful recovery of label permutation matrix (i.e. $\hat{\Pi} = \Pi^\natural$) means that inequality

$$\begin{aligned} & \langle -\psi(\Pi X B^\natural) + Y \circ \Pi X B^\natural \rangle \\ < \quad & \langle -\psi(\Pi^\natural X B^\natural) + Y \circ \Pi^\natural X B^\natural \rangle \end{aligned}$$

holds for any $\Pi \neq \Pi^\natural$. In other words, we need to identify certain sufficient conditions to ensure that the following probability

$$\begin{aligned} & P\big(\sup_{\Pi \neq \Pi^\natural} \langle -\psi(\Pi X B) + Y \circ \Pi X B \rangle \\ & \geq \langle -\psi(\Pi^\natural X B) + Y \circ \Pi^\natural X B \rangle \big) \end{aligned}$$

is vanishing as both n and m go to infinity in a suitable asymptotic regime.

Before moving to our main results, we first introduce the following row-wise quantities.

Information Gap For each pair of individuals i and j, we define

$$\Delta_{ij} := \langle \psi'(\boldsymbol{\lambda}_i) \circ \boldsymbol{\lambda}_i - \psi(\boldsymbol{\lambda}_i) \rangle - \langle \psi'(\boldsymbol{\lambda}_i) \circ \boldsymbol{\lambda}_j - \psi(\boldsymbol{\lambda}_j) \rangle, \quad (4)$$

where $\boldsymbol{\lambda}_i = (\mathbf{x}_i^T \mathbf{b}_1^\natural, \ldots, \mathbf{x}_i^T \mathbf{b}_m^\natural)$ for $i \in [n]$.

Variance For each pair of individuals i and j, we define

$$\begin{aligned} v_{ij} & := \langle \psi''(\boldsymbol{\lambda}_i) \circ (\boldsymbol{\lambda}_i - \boldsymbol{\lambda}_j)^2 \rangle \\ & = \sum_{l=1}^m \psi''(\boldsymbol{\lambda}_i[l])(\boldsymbol{\lambda}_i[l] - \boldsymbol{\lambda}_j[l])^2, \end{aligned} \quad (5)$$

where ψ' and ψ'' are the first and second order derivative of function ψ.

It can be checked that Δ_{ij} and v_{ij} are the expectation and variance of $\langle \mathbf{y}_i \circ \boldsymbol{\lambda}_i - \psi(\boldsymbol{\lambda}_i) \rangle - \langle \mathbf{y}_i \circ \boldsymbol{\lambda}_j - \psi(\boldsymbol{\lambda}_j) \rangle$, respectively. We can show that all labels are distinguishable when Δ_{ij} is relatively large compared with v_{ij} for all pairs of i and j. In other words, $\min_{i,j \in [n]} \Delta_{ij}^2 / v_{ij}$ can be viewed as the counterpart of *signal-to-noise ratio* (SNR, Pananjady et al. (2018)) in the linear models.

Theorem 3.1. *Assume $B^\natural$ is known and suppose X, $B^\natural$ and $\Pi^\natural$ satisfy that*

$$\Delta_{ij} \gtrsim \sqrt{(\log n) v_{ij}} \text{ and } v_{ij} \gtrsim \log n \quad \forall i, j \in [n]. \quad (6)$$

We have the perfect label recovery with high probability,

$$\begin{aligned} & P(\hat{\Pi} \neq \Pi^\natural) \\ \leq \quad & n^2 \max_{i \neq j} \max\{\exp\{-\Delta_{ij}^2/(4v_{ij})\}, \exp\{-v_{ij} c_\psi^2/4\}\} \\ \to \quad & 0, \end{aligned}$$

where c_ψ is a constant depending on function ψ.

When each element of X are i.i.d. sub-Gaussian random variables, it can be checked that both $\Delta_{ij} = \Theta(m)$ and $v_{ij} = \Theta(m)$. By simplifying requirements in (6), it suffices to have $m \gtrsim \log n$ to satisfy (6).

3.2. Examples

In this section, multiple examples are given to illustrate the relationship between m and n (see Section D for more detailed explanations) for perfect permutation recovery. Examples 3.2 - 3.5 given below describe four different data generation mechanisms. We can observe that, in those cases, it is impossible to recover $\Pi^\natural$ when $m = 1$. Hence, the sufficient conditions,

$$\min_{i_1 \neq i_2} |\psi'(\lambda_{i_1}) - \psi'(\lambda_{i_2})| \gtrsim \sqrt{\log n} \text{ (linear model)},$$

$$\min_{i_1 \neq i_2} |\sqrt{\psi'(\lambda_{i_1})} - \sqrt{\psi'(\lambda_{i_2})}| \gtrsim \sqrt{\log n} \text{ (poisson model)}$$

given by Wang et al. (2020) are too restrictive.

Example 3.2. Consider the scenario $p = 1$, then $X = \mathbf{x}$ is a vector. Assume $\mathbf{x}[i] \sim \text{Uniform}[a_1, a_2]$ for all i and assume each entry of $B^\natural$ is bounded between b_1 and b_2 ($0 < b_1 < b_2$). Without loss of generality, a_1, a_2, b_1, b_2 are all positive. We define $x_{gap} := \min_{i,j} x_{gap,ij} := \min_{i,j} |\mathbf{x}[i] - \mathbf{x}[j]|$, which is the minimum difference between any pair, $\mathbf{x}[i]$ and $\mathbf{x}[j]$. It is not hard to see that $x_{gap} = \Theta_p(\frac{1}{n^2})$. Moreover, when $m \gtrsim n^4 \log n$, it is sufficient for recovery of $\Pi^\natural$.

Example 3.3. Consider the scenario $p = log_2(n) + 1$ and $n = 2^{n_1}$ (n_1 is a positive integer) and the design matrix X

is complete in the sense that it satisfies

$$X = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & 1 \end{pmatrix}.$$

For instance, in the educational testing (Templin and Henson, 2010), each of n rows corresponds to a student with certain skill sets. The first column represents the intercept and rest of columns represent p different skills. Entries are binary-valued indicating the possess (1) or non-possess (0) of certain skills for different students. Without loss of generality, we assume that each entry of $B^\sharp$ is generated from the standard normal distribution. In this case, it suffices to require $m \geq \log n$ for correctly estimating $\Pi^\sharp$ for any strictly convex ψ with bounded second derivative.

Example 3.4. Consider the scenario p and n with $C_p^s \geq n$. (C_p^s is the sth coefficient in polynomial $(1+x)^p$) with s being a fixed constant. The design matrix X is sparse and bounded, that is, X satisfies that $\|\mathbf{x}_i\|_0 \leq s$ for any $i \in [n]$ and $a_1 \leq |X| \leq a_2$. Each entry of matrix $B^\sharp$ is generated by a standard normal distribution. We also assume that each row of X has different support. Under this setting, $m \gtrsim \log n$ suffices for permutation recovery for any strictly convex ψ with bounded second derivative.

Example 3.5. Consider the scenario that each entry of X follows $N(0, 1/p)$ independently and entry of $B^\sharp$ is generated from $N(0,1)$ independently. Under this setting, it can be shown that $m \gtrsim \log n$ suffices for permutation recovery for any strictly convex ψ with bounded second derivative.

3.3. On the lower bound of m

In this section, we discuss the minimum required number of m for permutation recovery. In particular, it suffices to consider the case that $B^\sharp$ is known, which is an easier task compared with the case when $B^\sharp$ is unknown. To start with, we recall the following Fano's lemma (Assouad, 1996).

Lemma 3.6 (Fano's Lemma)**.** *Let X be a random variable following probability distribution f, where f is from set $\{f_1, \ldots, f_{r+1}\}$ which satisfies that*

$$KL(f_i \| f_j) \leq \beta \ \text{for all} \ \ i \neq j.$$

Let $\psi(X) \in \{1, \ldots, r+1\}$ be an estimate of index of distribution. Then

$$\inf_{\psi} \sup_i P(\psi(X) \neq i) \geq 1 - \frac{\beta + \log 2}{\log r}. \tag{7}$$

We consider the following $n!$ models. For any permutation matrix Π_k ($k = 1, \ldots, n!$), we define f_k as the probability distribution of $Y = \Pi_k Y^\sharp$. Therefore, the KL divergence between f_{k_1} and f_{k_2} is

$$\begin{aligned} & KL(f_{k_1} \| f_{k_2}) \\ =& \mathbf{E}_{f_{k_1}} \langle Y \circ \mathbf{\Lambda}_{k_1} - \psi(\mathbf{\Lambda}_{k_1}) \rangle - \mathbf{E}_{f_{k_1}} \langle Y \circ \mathbf{\Lambda}_{k_2} - \psi(\mathbf{\Lambda}_{k_2}) \rangle \\ :=& \Lambda_{k_1}(\Pi_{k_1}, B^\sharp) - \Lambda_{k_1}(\Pi_{k_2}, B^\sharp), \end{aligned} \tag{8}$$

where $\mathbf{\Lambda}_k := \Pi_k X B^\sharp$. We let $\Delta(X, B^\sharp) := \max_{k_1 \neq k_2} \{\Lambda_{k_1}(\Pi_{k_1}, B^\sharp) - \Lambda_{k_1}(\Pi_{k_2}, B^\sharp)\}$. Therefore, by Fano's lemma, we have

$$\inf_{\hat{\Pi}} \sup_{\Pi_k} P(\hat{\Pi} \neq \Pi_k) \geq 1 - \frac{\Delta(X, B^\sharp) + \log 2}{\log(n!)}.$$

In particular, if $\Delta(X, B^\sharp) \leq Cmn$, we consequently have

$$\inf_{\hat{\Pi}} \sup_{\Pi_k} P(\hat{\Pi} \neq \Pi_k) \geq 1 - \frac{Cmn + \log 2}{n \log n} \geq 1/2, \tag{9}$$

when $m \lesssim \log n$. In other words, $m = \log n$ is the minimal number for perfect permutation recovery up to a multiplicative constant in Example 3.5.

On the other hand, $m \gtrsim \log n$ may not be tight under some situations. For instance, in Example 1, we can see that there exist i_1 and i_2 such that $|\mathbf{x}_{i_1} - \mathbf{x}_{i_2}|$ is $\Theta(1/n^2)$. We let $\Pi_1 = \mathbf{I}$ and set $\Pi_2(i) = i$ for $i \neq i_1, i_2$ and $\Pi_2(i_1) = i_2$ and $\Pi_2(i_2) = i_1$. Under such case, we have the following result.

Corollary 3.7. *By the constructions of Π_1 and Π_2, we have*

$$\inf_{\hat{\Pi}} \sup_{\Pi_k \in \{\Pi_1, \Pi_2\}} P(\hat{\Pi} \neq \Pi_k) \geq 1 - \frac{cm/n^4 + \log 2}{\log 2}. \tag{10}$$

Thus, the minimum requirement of m is at least of order n^4 ($\gg 1$) in Example 3.2.

4. Recovery Analysis when $B^\sharp$ is unknown

However, in practice, we have no prior knowledge of B and need to estimate the parameter matrix B and permutation matrix Π simultaneously. For a fixed Π, we define $\hat{B}(\Pi)$ to be the best estimator maximizing the log-likelihood function, that is,

$$\hat{B}(\Pi) = \arg\max_B \langle -\psi(\Pi X B) + Y \circ \Pi X B \rangle. \tag{11}$$

On the computational side, this is a concave optimization and $\hat{B}(\Pi)$ can be solved efficiently. On theoretical side, $\hat{B}(\Pi)$ does not admit explicit form which makes analysis harder. In the following, we discuss situations under which the labels can be recovered perfectly.

First of all, we note that the model is not identifiable when $p \geq n$ and X has full row rank. This is because, there exists a p by n matrix P_x such that $I = XP_x$. We can find that

$$\Pi^\sharp X B^\sharp = \Pi^\sharp \Pi \Pi X B^\sharp = (\Pi^\sharp \Pi) X (P_x \Pi X B^\sharp).$$

Thus, the underlying $\Pi^\sharp$ and $B^\sharp$ are again not identifiable. Such non-identifiability means that we have no chance to recover the true label permutation matrix, since there exist multiple global optimal values. In Unnikrishnan et al. (2015), it is further shown that $n \geq 2p$ is a necessary condition for perfect permutation recovery under the linear model setting with $m = 1$ and zero noise.

In the rest of paper, we only consider the case that $p < n$ for the generalized linear model. Moreover, we only focus on the situation that p is n^a with $a < 1/2$ so that the estimator has nice asymptotic properties even if we do not know the true $\Pi^\sharp$.

We recall the definition of log-likelihood function $L(\Pi, B) = \langle -\psi(\Pi X B) + Y \circ (\Pi X B)\rangle$ and introduce its corresponding population version,

$$\begin{aligned}\Lambda(\Pi, B) :=& \mathbb{E}\langle -\psi(\Pi X B) + Y \circ (\Pi X B)\rangle \\ =& \langle -\psi(\Pi X B) + \psi'(\Pi^\sharp X B^\sharp) \circ (\Pi X B)\rangle,\end{aligned} \tag{12}$$

where the expectation is taken with respect to Y.

4.1. Scenario 1: $d(\mathbf{I}, \Pi^\sharp)$ is small

If we have the prior knowledge that the underlying permutation matrix $\Pi^\sharp$ is close to $\mathbf{I}$ (i.e. $d(\mathbf{I}, \Pi^\sharp)$ is small), we then consider a "two-step estimation" computational method for dealing such case. In the first step, "two-step" algorithm aims to find a reasonable estimator of $B^\sharp$ by treating $\Pi = \mathbf{I}$. In the second step, we plug in this estimator to the objective and obtain the permutation matrix by maximizing the log-likelihood function. The implementation details are given in Algorithm 1.

Algorithm 1 Two-step Estimation.

Input: Observation matrix Y and design matrix X.
Output: Estimated permutation matrix $\hat{\Pi}$ and estimated coefficient matrix $\hat{B}$.
1. Solve $\hat{B} := \arg\max_B \{\langle -\psi(XB) + Y \circ XB\rangle\}$.
2. Solve $\hat{\Pi} := \arg\max_\Pi \{\langle -\psi(\Pi X \hat{B}) + Y \circ \Pi X \hat{B}\rangle\}$.

In below, we provide a theoretical analysis of the proposed two-step estimator. Under mild conditions, we show that $\hat{\Pi}$ returned by Algorithm 1 perfectly matches $\Pi^\sharp$ with high probability. To start with, we first assume the following assumptions on function ψ and design matrix, X.

A0 We assume that $\psi''(\cdot)$ is either monotonic or bounded.

A1 Each entry of X is bounded (i.e. $|X[i,j]| \leq C_0$ for universal constant C_0).

A2 There exist constants $c_1 > 0$, and γ_{1p} (which may depend on p) such that $\sharp\{i : X[i,:]\mathbf{b} \geq c_1\} \geq n/\gamma_{1p}$ and $\sharp\{i : X[i,:]\mathbf{b} \leq -c_1\} \geq n/\gamma_{1p}$ hold for any $\mathbf{b}$ with $\|\mathbf{b}\| = 1$.

Remark 4.1. Assumption *A0* is satisfied by most generalized linear models. For examples, ψ'' is bounded for Gaussian or Bernoulli distribution; ψ'' is monotonic for Poisson or Gamma distribution.

Remark 4.2. For a general n by p matrix X, its largest singular value is bounded by $\sqrt{n}\sqrt{p}\max_{i,j}|X[i,j]| = O(\sqrt{np})$. For an $n \times p$ matrix X with each entry being sampled from sub-Gaussian distribution, then its largest singular value is $O_p(\sqrt{n} + \sqrt{p})$. (Here we say Z is a sub-Gaussian random variable if $\mathbb{E}\exp\{tZ\} \leq \exp\{t^2\sigma^2/2\}$ holds for all $t > 0$ and fixed constant σ.)

Remark 4.3. It can be checked that Assumption *A2* is satisfied with high probability when X is a matrix with i.i.d sub-Gaussian random variables as its entries. Under such case, γ_{1p} is reduced to some constant. Furthermore, *A2* tells us that the smallest singular value of X is bounded from below. In fact, $\sigma_{min}(X) \geq \sqrt{c_1^2 n/\gamma_{1p}} = c_1\sqrt{n}/\sqrt{\gamma_{1p}}$.

We further need introduce the following notations: $x_{max} := \max_i \|\mathbf{x}_i\|$, $\psi'^\sharp_{max} := \max_{i,j}\psi'(\mathbf{x}_i^T\mathbf{b}_j^\sharp)$, $\psi''^\sharp_{max} := \max_{i,j}\psi'(\mathbf{x}_i^T\mathbf{b}_j^\sharp) \vee 1$, and $\psi''^\sharp_{min} := \min_{i,j}\psi''(\mathbf{x}_i^T\mathbf{b}_j^\sharp)$ which represent the maximum expected value of Y_{ij}'s, maximum variance/minimum variance of Y_{ij}'s respectively. We also write $\psi^\sharp_{cb} = \psi'^\sharp_{max} + \psi''^\sharp_{max}$ and define permutation-wise variance term, $v_{\Pi,partial}$

$$= \sum_{i:\Pi^\sharp(i)\neq\Pi(i)} \sum_{l=1}^m \psi''(\boldsymbol{\lambda}^\sharp_{\Pi^\sharp(i)}[l])(\boldsymbol{\lambda}^\sharp_{\Pi(i)}[l] - \boldsymbol{\lambda}^\sharp_{\Pi^\sharp(i)}[l])^2$$

and minimum pairwise variance term v_{min}

$$= \min_{i,j} \sum_{l=1}^m \psi''(\boldsymbol{\lambda}^\sharp_i[l])(\boldsymbol{\lambda}^\sharp_i[l] - \boldsymbol{\lambda}^\sharp_j[l])^2$$

to quantify the differences between X's rows.

Theorem 4.4. *With the knowledge that $d(\mathbf{I}, \Pi^\sharp) \leq h_{max}$ and assumptions A0 - A2, we also assume that $p = O(n^a)$ ($a < \frac{1}{2}$) and $h_{max} \lesssim n/(p\gamma_{1p}\log n)$. Then it holds that*

$$\|\hat{\mathbf{b}}_l - \mathbf{b}^\sharp_l\| \tag{13}$$

$$= O_p(\underbrace{\frac{\sqrt{p}(\sqrt{\psi^\sharp_{cb}}\sqrt{n - h_{max}} + \psi^\sharp_{cb} h_{max}\log n)}{n\psi''^\sharp_{min}}\gamma_{1p}}_{=:\delta^*})$$

for $l \in [m]$. Furthermore, if

$$\Lambda(\Pi^\sharp, B^\sharp) - \Lambda(\Pi, B^\sharp) \gtrsim v_{\Pi,partial} \cdot \sqrt{\log n / v_{min}} \quad (14)$$

and

$$\Lambda(\Pi^\sharp, B^\sharp) - \Lambda(\Pi, B^\sharp) \gtrsim md(\Pi, \Pi^\sharp)\psi^\sharp_{cb} x_{max} \delta^*, \quad (15)$$

then it holds that $P(\hat{\Pi} \neq \Pi^\sharp) \to 0$.

Here, δ^* defined in (13) is the estimation error for regression parameter. The first term can be viewed as the variance term and the second term is the bias term. Conditions (14) and (15) require that the information gap should dominate the errors caused by variance and bias, respectively. The requirement $h_{max} \leq n/(p\gamma_{1p} \log n)$ restricts the number of permuted labels. When p is fixed, then $n/(\log n)$ is similar to the condition $h_{max} \leq cn/(\log(n/h_{max}))$ required in Slawski et al. (2020) for linear models. The additional p in the denominator appears since that we do not have the access to the explicit form of $\hat{B}(\Pi)$ in generalized linear model setting.

Particularly, when X is sub-Gaussian, we can find that $\Lambda(\Pi^\sharp, B^\sharp) - \Lambda(\Pi, B^\sharp)$ is $\Omega(md(\Pi, \Pi^\sharp))$, $v_{\Pi,partial}$ is $O(md(\Pi, \Pi^\sharp))$ and $v_{min} = \Omega(m)$ when $\boldsymbol{\lambda}^\sharp_i[l]$ is bounded and $\min_{i,j} \|\boldsymbol{\lambda}^\sharp_i - \boldsymbol{\lambda}^\sharp_j\|_1$ is $\Omega(m)$. Then condition (14) can be simplified to

$$md(\Pi, \Pi^\sharp) \gtrsim md(\Pi, \Pi^\sharp)\sqrt{\log n / m},$$

which requires $m \geq K \log n$ for some large constant K. This leads to the following corollary.

Corollary 4.5. *Under the sub-Gaussian design matrix X with knowledge that $d(\mathbf{I}, \Pi^\sharp) \leq h_{max}$, we assume $h_{max} p = n/\log n$, $p = O(n^a)$ ($a < \frac{1}{2}$), and $m \gtrsim \log n$. We have*

$$P(\hat{\Pi} \neq \Pi^\sharp) \to 0 \quad \text{as } n \to \infty.$$

4.2. Scenario 2: no knowledge of $d(\mathbf{I}, \Pi^\sharp)$

For general $\Pi^\sharp$, we also aim to recover the underlying permutation matrix without any knowledge of $d(\mathbf{I}, \Pi^\sharp)$ and $B^\sharp$. To facilitate the theoretical analysis, we first define $B(\Pi) := \arg\max_B \Lambda(\Pi, B)$. Then it is straightforward to check that $B(\Pi^\sharp) = B^\sharp$. We also define

$$\Lambda(\Pi) := \max_B \Lambda(\Pi, B) = \Lambda(\Pi, B(\Pi)). \quad (16)$$

We additionally introduce the following permutation-wise quantities.

Information Gap For each permutation Π, we define

$$\Delta(X, B^\sharp, \Pi^\sharp, \Pi) = \Lambda(\Pi^\sharp) - \Lambda(\Pi). \quad (17)$$

This quantity can be interpreted as the information gap between permutation $\Pi^\sharp$ and Π. It can be seen that $\Delta(X, B^\sharp, \Pi^\sharp, \Pi)$ is always non-negative.

Variance For each fixed permutation Π, we define

$$v_{\Pi,B} = \sum_{i=1}^n \sum_{l=1}^m \psi''(\boldsymbol{\lambda}_{\Pi^\sharp(i)}[l])(\boldsymbol{\lambda}_{\Pi(i)}[l])^2. \quad (18)$$

It can be easily checked that $v_{\Pi,B}$ is the variance of $L(\Pi, B)$.

Theorem 4.6. *Under assumptions A0 - A2 and with no knowledge of $B^\sharp$ and $\Pi^\sharp$, we assume that $p = O(n^a)$ ($a < \frac{1}{2}$) and there exists a constant c_0 such that*

$$\Delta(X, B^\sharp, \Pi^\sharp, \Pi) \gtrsim \max\{\sqrt{(n+mp)mn\psi''^\sharp_{max} x^2_{max} \log n}, (n \log n + mp)x_{max}\} \quad (19)$$

for any Π satisfying $d(\Pi, \Pi^\sharp) > c_0 \frac{n}{p\gamma_{1p} \log n}$, and

$$\Lambda(\Pi^\sharp, B^\sharp) - \Lambda(\Pi, B^\sharp) \gtrsim \max\{v_{\Pi,partial} \cdot \sqrt{\log n / v_{min}}, md(\Pi, \Pi^\sharp)\psi^\sharp_{cb} x_{max} \delta^*\} \quad (20)$$

for any Π satisfying $d(\Pi, \Pi^\sharp) \leq c_0 \frac{n}{p\gamma_{1p} \log n}$. Then it holds that

$$P(\hat{\Pi} \neq \Pi^\sharp) \to 0 \quad (21)$$

as $n \to \infty$. Furthermore, $\|\hat{\mathbf{b}}_l - \mathbf{b}^\sharp_l\| = O_p(\frac{\gamma_{1p}\sqrt{p\psi^\sharp_{cb}}}{\sqrt{n}\psi^\sharp_{min}})$ for all $l \in [m]$.

Based on Theorem 4.6, we have the following corollary.

Corollary 4.7. *Without any knowledge of $B^\sharp$ and $\Pi^\sharp$, we assume that $\Delta(X, B^\sharp, \Pi^\sharp, \Pi) \geq c_1 md(\Pi, \Pi^\sharp)$ for $d(\Pi, \Pi^\sharp) > \frac{c_0 n}{p \log n}$ and $\Lambda(\Pi^\sharp, B^\sharp) - \Lambda(\Pi, B^\sharp) \geq c_2 md(\Pi, \Pi^\sharp)$ for $d(\Pi, \Pi^\sharp) \leq \frac{c_0 n}{p \log n}$ (c_0, c_1, c_2 are universal constants). Then it holds that $P(\hat{\Pi} \neq \Pi^\sharp) \to 0$, as long as $m/(x^2_{max} \log n) \to \infty$, $\gamma_{1p} = O(1)$ and $p = n^a$ ($0 < a < \frac{1}{2}$).*

4.3. On Computation of Maximum Likelihood Estimator

We aim to compute the maximum likelihood estimator,

$$(\hat{\Pi}, \hat{B}) = \arg\max_{\Pi, B} \langle -\psi(\Pi X B) + Y \circ \Pi X B \rangle. \quad (22)$$

Unfortunately, when $p > 1$, the above optimization problem (even for the linear models) is NP-hard. We consider a coordinate ascent method, i.e., alternatively maximizing Π given B and maximizing B given Π. The method will always converge to some stationary point, $\tilde{\Pi}, \tilde{B}$, satisfying $L(\tilde{\Pi}, \tilde{B}) \geq L(\Pi', \tilde{B})$ for $\forall\, \Pi' \neq \tilde{\Pi}$ and $\nabla_B L(\tilde{\Pi}, \tilde{B}) = \mathbf{0}$.

For the choice of a good initial permutation matrix Π_{ini}, we consider the following heuristic objective,

$$\Pi_{ini} = \arg\max_{\Pi} \langle \Pi, Y_\psi Y_\psi^T X X^T \rangle, \quad (23)$$

where $Y_\psi = (\psi')^{-1}(Y+1)$ be the inverse transformation of the original data through link function ψ'. The intuition is that $Y_\psi \approx \Pi^\natural X B^\natural$. It can be calculated that

$$\langle \Pi, Y_\psi Y_\psi^T X X^T \rangle \approx \langle \Pi, \Pi^\natural X B^\natural (\Pi^\natural X B^\natural)^T X X^T \rangle = \|B^\natural\|^2 \langle X, \Pi^\natural X \rangle \langle \Pi X, \Pi^\natural X \rangle, \quad (24)$$

when $m = 1$ and $p = 1$. If term $\langle X, \Pi^\natural X \rangle$ is positive, then the maximum value of (24) is achieved when $\Pi = \Pi^\natural$. Such warm start computational scheme is presented in Algorithm 2 and maximum likelihood estimation scheme is given in Algorithm 3.

Algorithm 2 Warm start for maximum likelihood estimation.

Input: Response matrix Y and design matrix X
Output: A good initial permutation matrix Π_{ini}.
1. Compute the matrix $Y_\psi = (\psi')^{-1}(Y+1)$.
2. Compute the matrix $C = Y_\psi Y_\psi^T X X^T$.
3. Solve $\Pi_{ini} := \arg\max_\Pi \langle \Pi, C \rangle$.
4. Return Π_{ini} as the initial Π.

Algorithm 3 Maximum likelihood (ML) estimation.

Input: Response matrix Y, design matrix X and initial permutation matrix Π_{ini}.
Output: Estimated permutation matrix $\hat{\Pi}$ and estimated coefficient matrix $\hat{B}$.
Let $\hat{\Pi} = \Pi_{ini}$.
while the likelihood is not converged **do**
  1. Solve $\hat{B} := \arg\max_B \{\langle -\psi(\hat{\Pi} X B) + Y \circ \hat{\Pi} X B \rangle\}$.
  2. Solve $\hat{\Pi} := \arg\max_\Pi \{\langle -\psi(\Pi X \hat{B}) + Y \circ \Pi X \hat{B} \rangle\}$.
end while
Return $\hat{B}$ and $\hat{\Pi}$.

4.4. Remarks

The technical challenge in the generalized linear model setting lies in the facts that the second derivative of likelihood function is associated with $\psi''(\cdot)$ functions which are not bounded. To be more specific, the Hessian matrix can be written as

$$\nabla^2 L(\mathbf{I}, \mathbf{b}) = -X^T D X,$$

with $D = \mathrm{diag}(d_1, \ldots, d_n)$ and $d_i = \psi''(\mathbf{x}_i^T \mathbf{b})$ when $m = 1$. For example, $\psi''(x) = \exp\{x\}$ is obviously unbounded for Poisson model. Moreover, for any fixed Π, the maximizer,

$$\hat{\mathbf{b}}(\Pi) = \arg\max_{\mathbf{b}} L(\Pi, \mathbf{b}),$$

does not admit the closed form and it brings additional difficulty. By contrast, there are no such afore-mentioned issues in the standard linear models.

In Wang et al. (2020), their analysis *only* considers the case when $m = 1$, $d(\mathbf{I}, \Pi^\natural)$ is small and $B^\natural$ is known. Our results include the most general cases, i.e., we have no prior knowledge of $\Pi^\natural$ and $B^\natural$. Therefore, we need to take special care of uniform bound over all possible permutations in our analyses.

From the computational perspective, note that both "two-step estimation" and "ML" methods require computing $\hat{\Pi} := \arg\max_\Pi \{\langle -\psi(\Pi X \hat{B}) + Y \circ \Pi X \hat{B} \rangle\}$. Since the number of candidates for the permutation matrix is n!, we cannot directly solve this optimization problem. However, we can reformulate this problem into a linear assignment problem which can be solved efficiently by specialized techniques such as Hungarian algorithm (Kuhn, 1955) or the Auction algorithm (Bertsekas and Castañón, 1992). To be more specific, we define an n by n cost matrix $C = (C[i,j])$, with

$$C[i,j] = \langle \psi(\mathbf{x}_i B) - \mathbf{y}_j \circ (\mathbf{x}_i B) \rangle.$$

Note that permutation matrix Π has only one non-zero element "1" in each row and column. It is then equivalent to solve the assignment problem,

$$\min_\tau \sum_{i \in [n]} C[\tau(i), i],$$

where τ is an one-to-one mapping from $[n]$ to $[n]$. Hence $\hat{\Pi}$ can be solved via using Hungarian algorithm or Auction algorithm.

Compared with Wang et al. (2020), they adopt an ℓ_1 regularization framework for computing $\hat{\Pi}$ and $\hat{B}$, while our method does not. This is due to the fact that Wang et al. (2020) assume a sparsely mismatch regime that $d(\mathbf{I}, \Pi^\natural)$ is small. That is the reason they introduce ℓ_1 regularizer to encourage recovering a sparse permutation. By contrast, we do not necessarily need $d(\mathbf{I}, \Pi^\natural)$ to be small in our theory.

5. Extension to Missing Observation Case

In many real applications, the observations are usually not fully observed (Rubin, 1976; Allison, 2001; Little and Rubin, 2019). The data may be missing at random during the collection process. Therefore, in this section, we generalize our results to the situations when data are not fully observed. To be specific, we consider the following model

$$Y_{miss} = E \circ Y = E \circ (\Pi^\natural Y^\natural), \quad (25)$$

where E is a binary matrix such that "1" means the entry is observed and "0" means the entry is missing. The elements in E are independent Bernoulli(q) random variables and q $(0 < q < 1)$ is the observation rate.

The log-likelihood function with missing observations can be written as

$$L(\Pi, B, E) = \langle E \circ (Y \circ (\Pi X B) - \psi(\Pi X B))\rangle$$

and its expectation can be computed as

$$\Lambda(\Pi, B, q) = q \cdot \Lambda(\Pi, B), \tag{26}$$

where the expectation is taken over both E and Y. We also define

$$\Lambda(\Pi, q) := \max_{B} \Lambda(\Pi, B, q) = q \cdot \Lambda(\Pi).$$

5.1. When $B^\sharp$ is known

In this scenario, we similarly define the following terms, the row-wise information gap $\Delta_{ij}(q)$ and the row-wise variance $v_{ij}(q)$. (See appendix for exact definition.)

Theorem 5.1. *Assume $B^\sharp$ is known and suppose X, $B^\sharp$, $\Pi^\sharp$ satisfy that*

$$\Delta_{ij}(q) \gtrsim \sqrt{(\log n) v_{ij}(q)} \text{ and } v_{ij}(q) \gtrsim \log n \ \ \forall i, j \in [n],$$

then it holds that

$$\begin{aligned} & P(\hat{\Pi} \neq \Pi^\sharp) \\ \leq \ & n^2 \max_{i \neq j} \max\{\exp\{-\frac{\Delta_{ij}^2(q)}{8 v_{ij}(q)}\}, \exp\{-v_{ij}(q) c_\psi^2/8\}\}, \end{aligned}$$

for some constant c_ψ.

Remark 5.2. Especially when $\lambda_{il}^\sharp$'s are bounded and $\min_{i,j} \sum_{l \in [m]} (\lambda_{il}^\sharp - \lambda_{jl}^\sharp)^2 = \Omega(m)$, then $q \geq \frac{\log n}{m}$ is required for the perfect permutation recovery. In other words, the number of required observations (m) for each individual is reciprocal to observation rate (q).

5.2. When $B^\sharp$ is unknown and $d(\mathbf{I}, \Pi^\sharp)$ is small

Under this scenario, we further define the partial variance term $v_{\Pi, partial, q}$ and minimum variance gap $v_{min,q}$, whose detailed definitions are given in Appendix A. We also assume the following assumptions on design matrix.

E1. Entries of X are bounded by some constant C_0.

E2. Let $\mathcal{S}_l = \{i : E[i, l] = 1\}$ ($l = 1, \ldots, m$). There exist constants $c_2 > 0$ and γ_{2p} such that $\sharp\{i : \mathbf{x}_i^T \mathbf{b} \geq c_2, i \in \mathcal{S}_l\} \geq |\mathcal{S}_l|/\gamma_{2p}$ and $\sharp\{i : \mathbf{x}_i^T \mathbf{b} \leq -c_2, i \in \mathcal{S}_l\} \geq |\mathcal{S}_l|/\gamma_{2p}$ hold for any $\mathbf{b}$ with $\|\mathbf{b}\| = 1$.

Remark 5.3. Assumptions *E1* and *E2* are parallel to *A1* and *A2*. They put the restrictions on sub-design matrices $X[\mathcal{S}_l, :]$'s. In particular, *E2* holds by taking $\gamma_{2p} = \Theta(1)$ with high probability, when each entry of X follows i.i.d. standard Normal distribution and $qn \gtrsim p$.

Theorem 5.4. *With the knowledge that $d(\mathbf{I}, \Pi^\sharp) \leq h_{max}$ and assumptions A0, E1 and E2, we assume $p = O((qn)^a)$ ($a < \frac{1}{2}$) and $h_{max} \lesssim nq/(p\gamma_{2p} \log n)$. We then have that*

$$\|\hat{\mathbf{b}}_l - \mathbf{b}_l^\sharp\| = O_p(\delta_q^*), \tag{27}$$

where δ_q^* equals

$$\frac{\sqrt{q\psi_{max}^{''\sharp} + q(1-q)\psi_{cb}^{\sharp 2}}\sqrt{n - h_{max}} + \psi_{max}^{''\sharp} h_{max} \log n}{qn\psi_{min}^\sharp/(\gamma_{2p}\sqrt{p})}.$$

Furthermore, if

$$\Lambda(\Pi^\sharp, B^\sharp) - \Lambda(\Pi, B^\sharp) \gtrsim \frac{1}{q} v_{\Pi, partial, q} \cdot \sqrt{\log n / v_{min,q}}$$

and

$$\Lambda(\Pi^\sharp, B^\sharp) - \Lambda(\Pi, B^\sharp) \gtrsim \frac{1}{q} m d(\Pi, \Pi^\sharp) \psi_{cb}^\sharp x_{max} \delta_q^*,$$

then it holds that

$$P(\hat{\Pi} \neq \Pi^\sharp) \to 0.$$

Remark 5.5. Under the setting of sub-Gaussian design, it is sufficient to have $m \geq \log n / q$ for permutation recovery when $d(\mathbf{I}, \Pi^\sharp) \leq c_0 \frac{nq}{p \log n}$ for some constant c_0.

Remark 5.6. The permutation recovery result is also established, when there is no any prior knowledge of $B^\sharp$ and $d(I, \Pi^\sharp)$. See Appendix A for explanations.

6. Conclusion

In this paper, we provide theoretical analyses of label permutation problem for the generalized linear model. The theory takes multivariate responses into account and is established under three different scenarios, with knowledge of $B^\sharp$, with knowledge of $d(\mathbf{I}, \Pi^\sharp)$ and without any knowledge. Our results are more general and remove the stringent conditions which are required by the case when $m = 1$. A detailed comparisons with existing literature are also provided to emphasize the technical challenges of considered setting. We also propose two computational methods, "maximum likelihood estimation" algorithm with warm start and "two-step estimation" algorithm. When the proportion of permuted labels is not too large, both methods work effectively under different settings of generating design matrix X. We further extend our results to the missing observation setting which has never been considered in the literature of label permutation problem. Experimental results match our theoretical findings. In practice, our computational methods sometimes may fail to find the global optimum when the proportion of permuted labels is large. Developing more efficient estimation methods constitutes a further promising direction.

References

Abubakar Abid and James Zou. Stochastic EM for shuffled linear regression. *arXiv preprint arXiv:1804.00681*, 2018.

Abubakar Abid, Ada Poon, and James Zou. Linear regression with shuffled labels. *arXiv preprint arXiv:1705.01342*, 2017.

Paul D Allison. *Missing data*. Sage publications, 2001.

Bin Yu Assouad. Fano, and le cam. *Festschrift for Lucien Le Cam*, pages 423–435, 1996.

Zhidong Bai and Tailen Hsing. The broken sample problem. *Probability theory and related fields*, 131(4):528–552, 2005.

Fadoua Balabdaoui, Charles R. Doss, and Cécile Durot. Unlinked monotone regression. *J. Mach. Learn. Res.*, 22:172:1–172:60, 2021.

Dimitri P. Bertsekas and David A. Castañón. A forward/reverse auction algorithm for asymmetric assignment problems. *Comput. Optim. Appl.*, 1(3):277–297, 1992.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine learning*, 3 (1):1–122, 2011.

Alexandra Carpentier and Teresa Schlueter. Learning relationships between data obtained independently. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 658–666, Cadiz, Spain, 2016.

Hock-Peng Chan and Wei-Liem Loh. A file linkage problem of degroot and goel revisited. *Statistica Sinica*, pages 1031–1045, 2001.

Gianpaolo Cugola and Alessandro Margara. Processing flows of information: From data stream to complex event processing. *ACM Comput. Surv.*, 44(3):15:1–15:62, 2012.

Morris H DeGroot and Prem K Goel. The matching problem for multivariate normal data. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 14–29, 1976.

Morris H DeGroot and Prem K Goel. Estimation of the correlation coefficient from a broken random sample. *The Annals of Statistics*, 8(2):264–278, 1980.

Morris H DeGroot, Paul I Feder, and Prem K Goel. Matchmaking. *The Annals of Mathematical Statistics*, 42(2): 578–593, 1971.

Valentin Emiya, Antoine Bonnefoy, Laurent Daudet, and Rémi Gribonval. Compressed sensing with unknown sensor permutation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1040–1044, Florence, Italy, 2014.

Ivan P Fellegi and Alan B Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64 (328):1183–1210, 1969.

Nicolas Flammarion, Cheng Mao, and Philippe Rigollet. Optimal rates of statistical seriation. *Bernoulli*, 25(1): 623–653, 2019.

Thomas R Fleming and David P Harrington. *Counting processes and survival analysis*, volume 169. John Wiley & Sons, 2011.

Prem K Goel. On re-pairing observations in a broken random sample. *The Annals of Statistics*, 3(6):1364–1369, 1975.

Trevor Hastie, Rahul Mazumder, Jason D. Lee, and Reza Zadeh. Matrix completion and low-rank SVD via fast alternating least squares. *J. Mach. Learn. Res.*, 16:3367–3402, 2015.

Daniel J. Hsu, Kevin Shi, and Xiaorui Sun. Linear regression without correspondence. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1531–1540, Long Beach, CA, 2017.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

Rong Ma, T Tony Cai, and Hongzhe Li. Optimal permutation recovery in permuted monotone matrix model. *Journal of the American Statistical Association*, pages 1–15, 2020.

Ivan Nazarov, Boris Shirokikh, Maria Burkina, Gennady Fedonin, and Maxim Panov. Sparse group inductive matrix completion. *arXiv preprint arXiv:1804.10653*, 2018.

Howard B. Newcombe and James M. Kennedy. Record linkage: making maximum use of the discriminating power of identifying information. *Commun. ACM*, 5(11): 563–566, 1962.

Ashwin Pananjady, Martin J. Wainwright, and Thomas A. Courtade. Denoising linear models with permuted data. In *Proceedings of the 2017 IEEE International Symposium on Information Theory (ISIT)*, pages 446–450, Aachen, Germany, 2017.

Ashwin Pananjady, Martin J. Wainwright, and Thomas A. Courtade. Linear regression with shuffled data: Statistical and computational limits of permutation recovery. *IEEE Trans. Inf. Theory*, 64(5):3286–3300, 2018.

Liangzu Peng and Manolis C. Tsakiris. Linear regression without correspondences via concave minimization. *IEEE Signal Process. Lett.*, 27:1580–1584, 2020.

Philippe Rigollet and Jonathan Weed. Uncoupled isotonic regression via minimum wasserstein deconvolution. *Information and Inference: A Journal of the IMA*, 8(4): 691–717, 2019.

Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Xu Shi, Xiaoou Li, and Tianxi Cai. Spherical regression under mismatch corruption with application to automated knowledge translation. *Journal of the American Statistical Association*, pages 1–12, 2020.

Yandong Shi and Yuanming Shi. Learning to branch-and-bound for header-free communications. In *Proceedings of the 2019 IEEE Globecom Workshops*, pages 1–6, Waikoloa, HI, 2019.

Martin Slawski and Emanuel Ben-David. Linear regression with sparsely permuted data. *Electronic Journal of Statistics*, 13(1):1–36, 2019.

Martin Slawski, Mostafa Rahmani, and Ping Li. A sparse representation-based approach to linear regression with partially shuffled labels. In *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 38–48, Tel Aviv, Israel, 2019.

Martin Slawski, Emanuel Ben-David, and Ping Li. Two-stage approach to multivariate linear regression with sparsely mismatched data. *J. Mach. Learn. Res.*, 21: 204:1–204:42, 2020.

Martin Slawski, Guoqing Diao, and Emanuel Ben-David. A pseudo-likelihood approach to linear regression with partially shuffled data. *J. Comput. Graph. Stat.*, 30(4): 991–1003, 2021.

Jonathan Templin and Robert A Henson. *Diagnostic measurement: Theory, methods, and applications*. Guilford Press, 2010.

Manolis C Tsakiris. Eigenspace conditions for homomorphic sensing. *arXiv preprint arXiv:1812.07966*, 2018.

Manolis C Tsakiris, Liangzu Peng, Aldo Conca, Laurent Kneip, Yuanming Shi, and Hayoung Choi. An algebraic-geometric approach to shuffled linear regression. *arXiv preprint arXiv:1810.05440*, 2018.

Jayakrishnan Unnikrishnan, Saeid Haghighatshoar, and Martin Vetterli. Unlabeled sensing: Solving a linear system with unordered measurements. In *Proceedings of the 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 786–793, Monticello, IL, 2015.

Jayakrishnan Unnikrishnan, Saeid Haghighatshoar, and Martin Vetterli. Unlabeled sensing with random linear measurements. *IEEE Trans. Inf. Theory*, 64(5):3237–3253, 2018.

Guanyu Wang, Jiang Zhu, Rick S. Blum, Peter Willett, Stefano Maranò, Vincenzo Matta, and Paolo Braca. Signal amplitude estimation and detection from unlabeled binary quantized samples. *IEEE Trans. Signal Process.*, 66(16):4291–4303, 2018.

Zhenbang Wang, Emanuel Ben-David, and Martin Slawski. Estimation in exponential family regression based on linked data contaminated by mismatch error. *arXiv preprint arXiv:2010.00181*, 2020.

Hang Zhang and Ping Li. Optimal estimator for unlabeled linear regression. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 11153–11162, Virtual Event, 2020.

Hang Zhang and Ping Li. Sparse recovery with shuffled labels: Statistical limits and practical estimators. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, pages 1760–1765, Melbourne, Australia, 2021.

Hang Zhang and Ping Li. Greed is good: correspondence recovery for unlabeled linear regression. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence (UAI)*, Pittsburgh, PA, 2023a.

Hang Zhang and Ping Li. One-step estimator for permuted sparse recovery. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, Honolulu, HI, 2023b.

Hang Zhang, Martin Slawski, and Ping Li. Permutation recovery from multiple measurement vectors in unlabeled sensing. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, pages 1857–1861, Paris, France, 2019.

Hang Zhang, Martin Slawski, and Ping Li. The benefits of diversity: Permutation recovery in unlabeled sensing from multiple measurement vectors. *IEEE Trans. Inf. Theory*, 68(4):2509–2529, 2022.

Appendices of "Regression with Label Permutation in Generalized Linear Model"

The organization of appendix is given as followed. In Section A, we provide detailed explanations of permutation recovery results when there exist missing data. In Section B, simulation results are given to help readers to understand our theory better. A real data application is shown in Section C to illustrate the effectiveness of our proposed algorithm. Section D is dedicated to explain the relationship between m and n for perfect permutation recovery. An explanation of $\Delta(X, B^\sharp, \Pi^\sharp, \Pi)$ is provided in Section E. All technical proofs are collected in Sections F - I. Finally, a remark on computational approach for linear models is given in Section J.

A. Missing details under "Missing Observation Case"

In this section, we provide more details of the case when there exist missing observations. The model becomes

$$Y_{miss} = E \circ Y = E \circ (\Pi^\sharp Y^\sharp), \tag{28}$$

where E is a binary matrix such that "1" means the entry is observed and "0" means the entry is missing. The elements in E are independent Bernoulli(q) random variables and q $(0 < q < 1)$ is the observation rate.

The log-likelihood function with missing observations can be written as

$$\begin{aligned} L(\Pi, B, E) &= \{\sum_{i,l:E[i,l]=1} Y[i,l](\mathbf{x}_{\Pi(i)}^T \mathbf{b}_l) - \psi(\mathbf{x}_{\Pi(i)}^T \mathbf{b}_l)\} \\ &= \langle E \circ (Y \circ (\Pi X B) - \psi(\Pi X B))\rangle \end{aligned} \tag{29}$$

and its expectation can be computed as

$$\begin{aligned} \Lambda(\Pi, B, q) &= \mathbb{E}\{\sum_{i,l:E[i,l]=1} Y[i,l](\mathbf{x}_{\Pi(i)}^T \mathbf{b}_l) - \psi(\mathbf{x}_{\Pi(i)}^T \mathbf{b}_l)\} \\ &= q \cdot \Lambda(\Pi, B), \end{aligned} \tag{30}$$

where the expectation is taken over both E and Y. We also define

$$\Lambda(\Pi, q) := \max_B \Lambda(\Pi, B, q) = q \cdot \Lambda(\Pi).$$

A.1. When $B^\sharp$ is known

In this scenario, we can similarly define the following terms, the row-wise information gap,

$$\begin{aligned} \Delta_{ij}(q) &:= \mathbb{E}\sum_l \mathbf{1}\{E[\Pi^\sharp(i), l] = 1\}\{(\psi'(\boldsymbol{\lambda}_i^\sharp[l])\boldsymbol{\lambda}_i^\sharp[l] - \psi(\boldsymbol{\lambda}_i^\sharp[l])) - (\psi'(\boldsymbol{\lambda}_i^\sharp[l])\boldsymbol{\lambda}_j^\sharp[l] - \psi(\boldsymbol{\lambda}_j^\sharp[l]))\} \\ &= q\sum_l \{(\psi'(\boldsymbol{\lambda}_i^\sharp[l])\boldsymbol{\lambda}_i^\sharp[l] - \psi(\boldsymbol{\lambda}_i^\sharp[l])) - (\psi'(\boldsymbol{\lambda}_i^\sharp[l])\boldsymbol{\lambda}_j^\sharp[l] - \psi(\boldsymbol{\lambda}_j^\sharp[l]))\} = q\Delta_{ij} \end{aligned} \tag{31}$$

and the row-wise variance,

$$\begin{aligned} v_{ij}(q) &:= q\sum_{l=1}^m \psi''(\boldsymbol{\lambda}_i^\sharp[l])(\boldsymbol{\lambda}_i^\sharp[l] - \boldsymbol{\lambda}_i^\sharp[l])^2 \\ &\quad + q(1-q)\sum_{l=1}^m (\psi'(\boldsymbol{\lambda}_i^\sharp[l])(\boldsymbol{\lambda}_i^\sharp[l] - \boldsymbol{\lambda}_j^\sharp[l]) - (\psi(\boldsymbol{\lambda}_i^\sharp[l]) - \psi(\boldsymbol{\lambda}_j^\sharp[l])))^2, \end{aligned} \tag{32}$$

which is the variance of $\mathrm{var}(\langle E[\Pi^\sharp(i), :] \circ (\mathbf{y}_i \circ \boldsymbol{\lambda}_i^\sharp - \psi(\boldsymbol{\lambda}_i^\sharp))\rangle - \langle E[\Pi^\sharp(i), :] \circ (\mathbf{y}_i \circ \boldsymbol{\lambda}_j^\sharp - \psi(\boldsymbol{\lambda}_j^\sharp))\rangle)$.

Theorem A.1. *[Restatement of Theorem 5.1] Assume $B^\sharp$ is known and suppose X, $B^\sharp$, $\Pi^\sharp$ satisfy that*

$$\Delta_{ij}(q) \gtrsim \sqrt{(\log n) v_{ij}(q)} \text{ and } v_{ij}(q) \gtrsim \log n \;\; \forall i, j \in [n], \tag{33}$$

then it holds that

$$P(\hat{\Pi} \neq \Pi^\sharp) \leq n^2 \max_{i \neq j} \max\{\exp\{-\frac{\Delta_{ij}^2(q)}{8v_{ij}(q)}\}, \exp\{-v_{ij}(q)c_\psi^2/8\}\}, \tag{34}$$

for some constant c_ψ.

Remark A.2. Especially when $\lambda_{il}^\sharp$'s are bounded and $\min_{i,j} \sum_{l \in [m]} (\lambda_{il}^\sharp - \lambda_{jl}^\sharp)^2 = \Omega(m)$, then $q \geq \frac{\log n}{m}$ is required for the perfect permutation recovery. In other words, the number of required observations (m) for each individual is reciprocal to observation rate (q).

A.2. When $B^\sharp$ is unknown and $d(\mathbf{I}, \Pi^\sharp)$ is small

Under this scenario, we further define the partial variance term

$$\begin{aligned} v_{\Pi,partial,q} \quad &= \quad q \sum_{i:\Pi(i) \neq \Pi^\sharp(i)} \sum_{l=1}^{m} \Big\{ \psi''(\boldsymbol{\lambda}_i^\sharp[l])(\boldsymbol{\lambda}_{\Pi(i)}[l]^\sharp - \boldsymbol{\lambda}_i[l]^\sharp)^2 \\ &\quad + q(1-q)(\psi'(\boldsymbol{\lambda}_i[l]^\sharp)(\boldsymbol{\lambda}_{\Pi(i)}[l]^\sharp - \boldsymbol{\lambda}_i[l]^\sharp) - \psi(\boldsymbol{\lambda}_{\Pi(i)}^\sharp[l]) + \psi(\boldsymbol{\lambda}_i[l]^\sharp))^2 \Big\} \end{aligned}$$

and minimum variance gap

$$v_{min,q} = \min_{i,j} \{q \sum_{l=1}^{m} \psi''(\boldsymbol{\lambda}_i^\sharp[l])(\boldsymbol{\lambda}_i[l]^\sharp - \boldsymbol{\lambda}_j[l]^\sharp)^2 + q(1-q) \sum_{l=1}^{m} (\psi'(\boldsymbol{\lambda}_i^\sharp[l])(\boldsymbol{\lambda}_i[l]^\sharp - \boldsymbol{\lambda}_j[l]^\sharp) - \psi(\boldsymbol{\lambda}_i[l]^\sharp) + \psi(\boldsymbol{\lambda}_j[l]^\sharp))^2\}.$$

We also assume the following assumptions on design matrix.

E1 Entries of X are bounded by some constant C_0.

E2 Let $\mathcal{S}_l = \{i : E[i,l] = 1\}$ $(l = 1, \ldots, m)$. There exist constants $c_2 > 0$ and γ_{2p} such that $\sharp\{i : \mathbf{x}_i^T \mathbf{b} \geq c_2, i \in \mathcal{S}_l\} \geq |\mathcal{S}_l|/\gamma_{2p}$ and $\sharp\{i : \mathbf{x}_i^T \mathbf{b} \leq -c_2, i \in \mathcal{S}_l\} \geq |\mathcal{S}_l|/\gamma_{2p}$ hold for any $\mathbf{b}$ with $\|\mathbf{b}\| = 1$.

Remark A.3. Assumptions *E1* and *E2* are parallel to *A1* and *A2*. They put the restrictions on sub-design matrices $X[\mathcal{S}_l, :]$'s. In particular, *E2* holds by taking $\gamma_{2p} = \Theta(1)$ with high probability, when each entry of X follows i.i.d. standard Normal distribution and $qn \gtrsim p$.

Theorem A.4. *[Restatement of Theorem 5.4] With the knowledge that $d(\mathbf{I}, \Pi^\sharp) \leq h_{max}$ and assumptions A0, E1 and E2, we assume $p = O((qn)^a)$ $(a < \frac{1}{2})$ and $h_{max} \lesssim nq/(p\gamma_{2p} \log n)$. We then have that*

$$\|\hat{\mathbf{b}}_l - \mathbf{b}_l^\sharp\| = O(\underbrace{\frac{\sqrt{p}(\sqrt{q\psi_{max}^{''\sharp} + q(1-q)\psi_{cb}^{\sharp 2}}\sqrt{n - h_{max}} + \psi_{max}^{''\sharp} h_{max} \log n)}{qn\psi_{min}^\sharp/\gamma_{2p}}}_{=:\delta_q^*}). \tag{35}$$

Furthermore, if

$$\Lambda(\Pi^\sharp, B^\sharp) - \Lambda(\Pi, B^\sharp) \gtrsim \frac{1}{q} v_{\Pi,partial,q} \cdot \sqrt{\log n / v_{min,q}} \tag{36}$$

and

$$\Lambda(\Pi^\sharp, B^\sharp) - \Lambda(\Pi, B^\sharp) \gtrsim \frac{1}{q} md(\Pi, \Pi^\sharp)\psi_{cb}^\sharp x_{max} \delta_q^*, \tag{37}$$

then it holds that

$$P(\hat{\Pi} \neq \Pi^\sharp) \to 0. \tag{38}$$

Remark A.5. Under the setting of sub-Gaussian design, it is sufficient to have $m \geq \log n/q$ for permutation recovery when $d(\mathbf{I}, \Pi^\sharp) \leq c_0 \frac{nq}{p \log n}$ for some constant c_0.

A.3. Without knowledge of $B^\sharp$ and $d(\mathbf{I}, \Pi^\sharp)$

In this situation, we further assume the following conditions.

> *E2'* There exist constants $c_3 > 0$ and γ_{3p} such that $\sharp\{i : \mathbf{x}_i^T \mathbf{b} \geq c_3, i \in \mathcal{S}\} \geq qn/\gamma_{3p}$ and $\sharp\{i : \mathbf{x}_i^T \mathbf{b} \leq -c_3, i \in \mathcal{S}\} \geq qn/\gamma_{3p}$ hold for any $\mathbf{b}$ with $\|\mathbf{b}\| = 1$ and $\mathcal{S}$ with $|\mathcal{S}| \geq qn/2$. (It is a modified and stronger version of *E2*.)

Additionally, we let $\Delta_q(X, B^\sharp, \Pi, \Pi^\sharp) := \Lambda(\Pi^{''\sharp}, q) - \Lambda(\Pi, q)$ which is equal to $q\Delta(X, B^\sharp, \Pi, \Pi^\sharp)$, and define the following variance-related quantity,

$$V_2(q) = (qn\psi^\sharp_{max} + q(1-q)n\psi^{\sharp 2}_{cb})(\psi(x_{max}))^2.$$

Theorem A.6. *Under assumptions A0, E1 and E2', we assume that there exists c_0 such that*

$$\Delta_q(X, B^\sharp, \Pi, \Pi^\sharp) \gtrsim \max\{\sqrt{(n\log n + mp\log(n))mV_2(q)}, \psi^\sharp_{cb}(n\log^2 n + mp\log n)\} \tag{39}$$

holds for any Π with $d(\Pi, \Pi^\sharp) > c_0 \frac{nq}{p\gamma_{3p}\log n}$, and

$$\Lambda(\Pi^\sharp, B^\sharp) - \Lambda(\Pi, B^\sharp) \gtrsim \max\{\frac{1}{q}v_{\Pi,partial,q} \cdot \sqrt{\log n / v_{min,q}}, \frac{1}{q}md(\Pi, \Pi^\sharp)\psi^\sharp_{cb}x_{max}\delta^*_q\} \tag{40}$$

holds for any Π with $d(\Pi, \Pi^\sharp) \leq c_0 \frac{nq}{p\gamma_{3p}\log n}$. ($\delta^*_q$ is the same as defined in (35) in Theorem A.4.)

Then it holds that

$$P(\hat{\Pi} \neq \Pi^\sharp) \to 0 \tag{41}$$

as $n \to \infty$. Furthermore,

$$\|\hat{\mathbf{b}}_l - \mathbf{b}^\sharp_l\| = O_p(\frac{\sqrt{pn}(\sqrt{q\psi^{''\sharp}_{max} + q(1-q)\psi^{\sharp 2}_{cb}})}{qn\psi^{''\sharp}_{min}/\gamma_{3p}}) \tag{42}$$

for all $l \in [m]$.

Especially, when γ_{3p}, $\psi^{''\sharp}_{min}$, $\psi^{''\sharp}_{max}$, $\psi^\sharp_{cb}$ are $O(1)$, and $\min_{i\neq j}\sum_{l\in[m]}|\boldsymbol{\lambda}^\sharp_i[l] - \boldsymbol{\lambda}^\sharp_j[l]| = \Omega(m)$, it suffices to require

$$m \gtrsim \frac{(\psi(x_{max}))^2 \log n}{q}, \; q \gtrsim \frac{p(\psi(x_{max}))^2\log n}{n}$$

for exact permutation recovery.

A.4. ML Estimator with Missing Observations

The warm-start stage of ML estimation algorithm is modified in the missing observation setting. In particular, we use SoftImpute (Hastie et al., 2015) method to impute the missing entries of the data matrix. The procedure is given as below in Algorithm 4.

Algorithm 4 ML estimation with warm start for missing observations.

Input: Observations with missing entries Y_{miss}, design matrix X
Output: A good initial permutation matrix Π_{ini}.
1. Compute the matrix $Y_{\psi,miss} = (\psi')^{-1}(Y_{miss} + 1)$.
2. Use SoftImpute method to complete the matrix, $Y_{\psi,miss}$, to get $Y_{\psi,comp}$.
3. Compute the matrix $C = Y_{\psi,comp}Y_{\psi,comp}^T XX^T$.
4. Solve $\Pi_{ini} := \arg\max_{\Pi}\langle \Pi, C\rangle$.
5. Set $\hat{\Pi} = \Pi_{ini}$ as the initial permutation matrix.
while The likelihood not converged **do**
  6.a. Solve $\hat{B} := \arg\max_B\{\langle -E \circ \psi(\hat{\Pi}XB) + E \circ Y_{miss} \circ \hat{\Pi}XB\rangle\}$.
  6.b. Solve $\hat{\Pi} := \arg\max_{\Pi}\{\langle -E \circ \psi(\Pi X\hat{B}) + E \circ Y_{miss} \circ \Pi X\hat{B}\rangle\}$.
end while
7. Return $\hat{B}$ and $\hat{\Pi}$.

A.5. Two-step Estimator with Missing Observations

Similarly, we introduce the two-step estimator under the missing observation setting. The procedure is given as follows in Algorithm 5.

Algorithm 5 Two-step Estimation with missing observations.

Input: Observations with missing entries Y_{miss}, indicator matrix E and design matrix X
Output: Estimated permutation matrix $\hat{\Pi}$ and estimated coefficient matrix $\hat{B}$.
1. Solve $\hat{B} := \arg\max_B\{\langle -E \circ \psi(XB) + E \circ Y_{miss} \circ XB\rangle\}$.
2. Solve $\hat{\Pi} := \arg\max_{\Pi}\{\langle -E \circ \psi(\Pi X\hat{B}) + E \circ Y_{miss} \circ \Pi X\hat{B}\rangle\}$.

B. Simulation Studies

Setting 1 In the first simulation setting, we consider to evaluate the performance of maximum likelihood estimation method. We set n to be 256 and 512 and let 25% or 33 % labels be permuted. We vary m from $\{\log_2 n, 2\log_2 n, \dots, 20\log_2 n\}$ and set observation rate q at different levels. For design matrix X, each row independently follows a multivariate Gaussian distribution $N(\mathbf{0}, I_p/p)$ ($p = 10$). For coefficient matrix B, each element is i.i.d. standard Gaussian random variable. The curves of probability for successful permutation recovery are plotted in Figure 1.

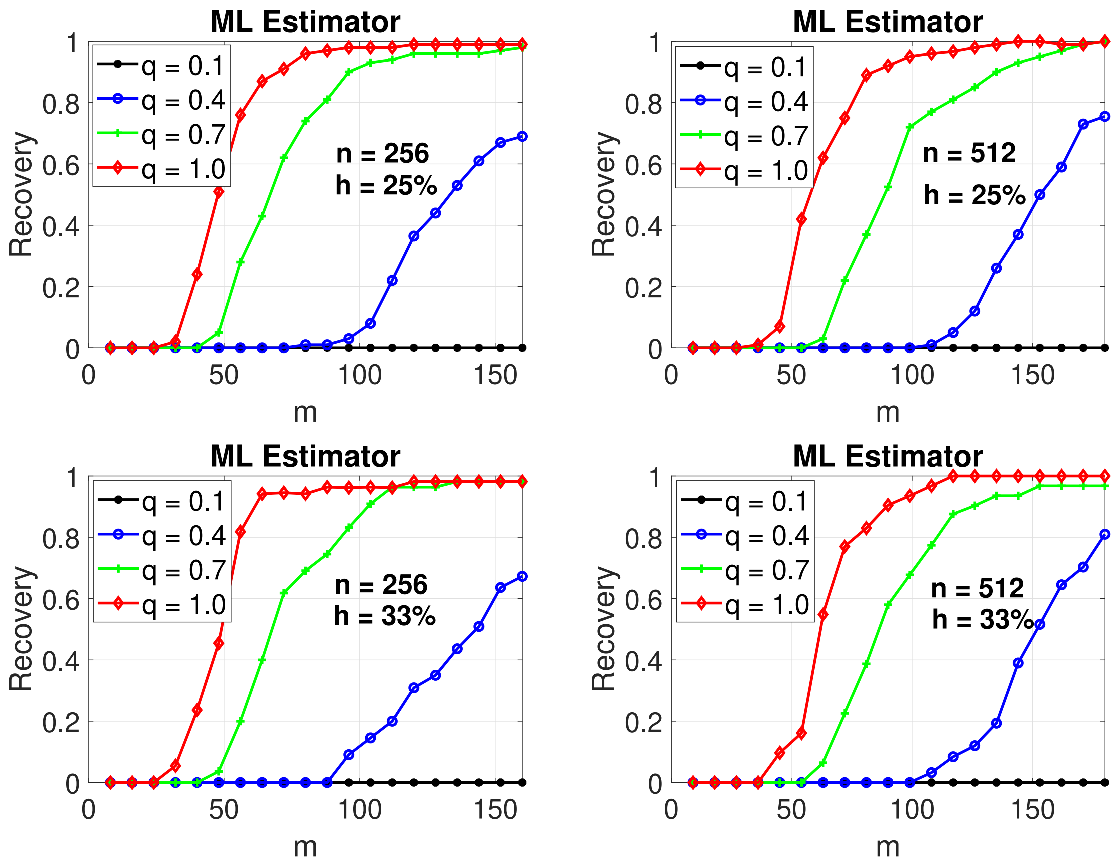

Figure 1. The curves of label permutation recovery under different m, q and h by using maximum likelihood estimation algorithm. Upper left: $n = 256, h = 0.25n$; Upper right: $n = 512, h = 0.25n$; Bottom left: $n = 256, h = 0.33n$; Bottom right: $n = 512, h = 0.33n$. Each point is the average of 500 replications.

Setting 2 In the second simulation setting, we illustrate the performance of two-step estimation method. We deliberately permute the true label by some proportions (5%, 10%, ..., 100 %). We set n to be 256 / 512 and set $m = 10\log_2 n$ / $20\log_2 n$. The observation rate q varies from 0.4 to 1.0. The design matrix and coefficient matrix remains the same as in the first setting. The curves of probability for successful permutation recovery are plotted in Figure 2.

Setting 3 In the third simulation setting, we compare with the results by fitting a linear model directly to the original data ("linear") or to the log-transform of data ("log-trans") under different generation schemes. (We use the ADMM-based algorithm in Section J for implementation.)

1. For design matrix X, each row independently follows a multivariate Gaussian distribution $N(\mathbf{0}, I_p/p)$. For coefficient matrix B, each element is i.i.d. standard Gaussian random variable. In this case, we set $n = 256$ and $p = 10$.

2. Matrix X is a complete design matrix . For coefficient matrix B, each element is i.i.d. uniform random variable on $U(0, 2)$. In this case, $n = 256$, $p = 1 + \log_2 n$.

3. For sparse design matrix X, each row has at most s non-zero entries and positions of non-zero elements are sampled uniformly. For coefficient matrix B, each element is i.i.d. uniform random variable on $U(0, 2)$. In this case, we set $n = 256$, $p = 20$ and $s = 5$.

Such comparisons under model mis-specification are shown in Figure 3.

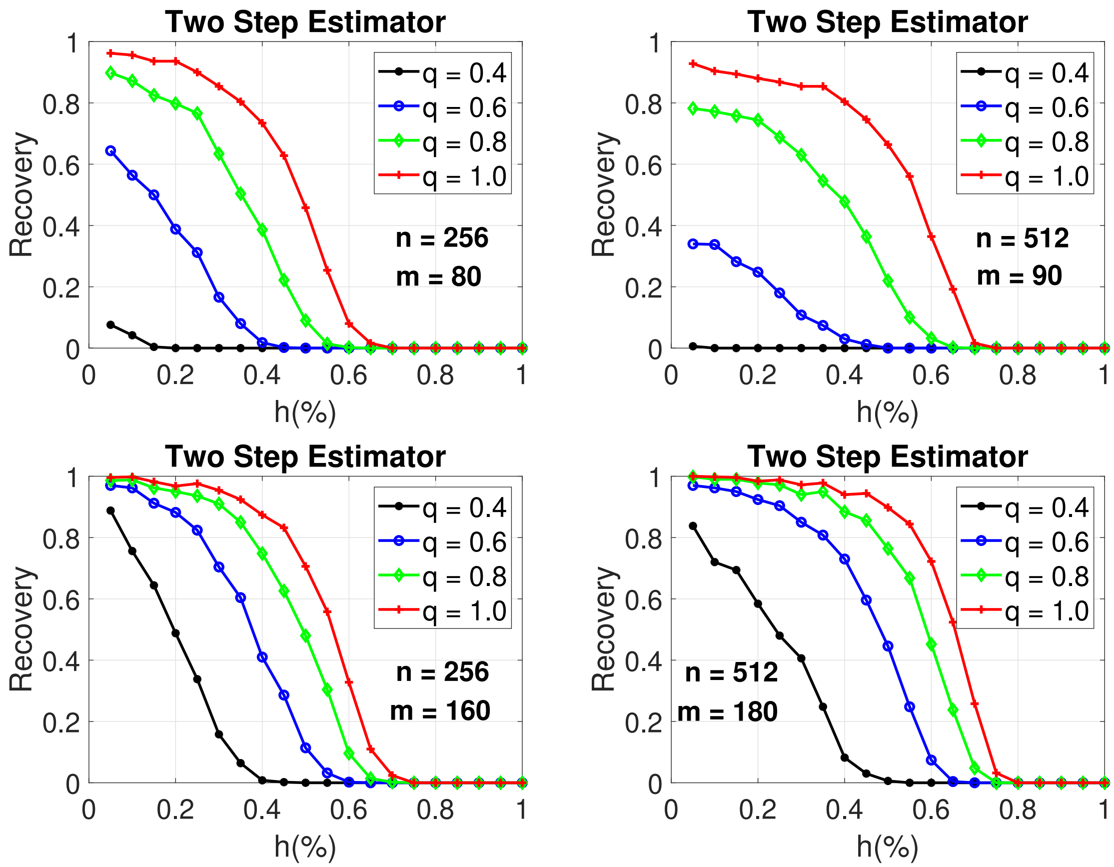

Figure 2. The curves of label permutation recovery under different n, m and h by using two-step estimation algorithm. Upper left: $n = 256, m = 80$; Upper right: $n = 512, m = 90$; Bottom left: $n = 256, m = 160$; Bottom Right: $n = 512, m = 180$. Each point is the average of 500 replications.

Setting 4 In the fourth simulation setting, we consider to evaluate the performance of maximum likelihood estimator when p varies. We set n to be 256 and 512 and let 25% of labels be permuted. We vary m from $\{\log_2 n, 2\log_2 n, \ldots, 20\log_2 n\}$. For design matrix X, each row independently follows a multivariate Gaussian distribution $N(\mathbf{0}, I_p/p)$ with $p = 5, 10, 15, 20$ or 25. For coefficient matrix B, each element is i.i.d. standard Gaussian random variable. The curves of probability for successful permutation recovery are shown in the bottom-right plot in Figure 3.

From Figure 1, we can see that the probability of successful label recovery increases as m increases. The probability changes drastically from 0 to 1 when $m \approx 10\log_2 n$. This matches our theory. In addition, we can see that m required for perfect permutation recovery gets larger as observation rate q decreases. From Figure 2, we can observe that the probability of successful label recovery decreases as proportion of wrong label increases. The probability changes drastically from 1 to 0 when 20% of individuals are given with wrong labels. Additionally, as the observation rate decreases, the successful recovery probability also decreases. From Figure 3, we can see that the recovery results will get worse if we fit the data generated from log-linear model by using linear methods. Thus, model mis-specification (i.e. non-Gaussian setting) may lead to bad recovery results. Furthermore, we can see that the value of p does not effect the recovery result when it is in the suitable regime of n, i.e., $p \gtrsim \log n$ and $p \lesssim n^{1/2}$. This matches our findings in theory and Example 3.5.

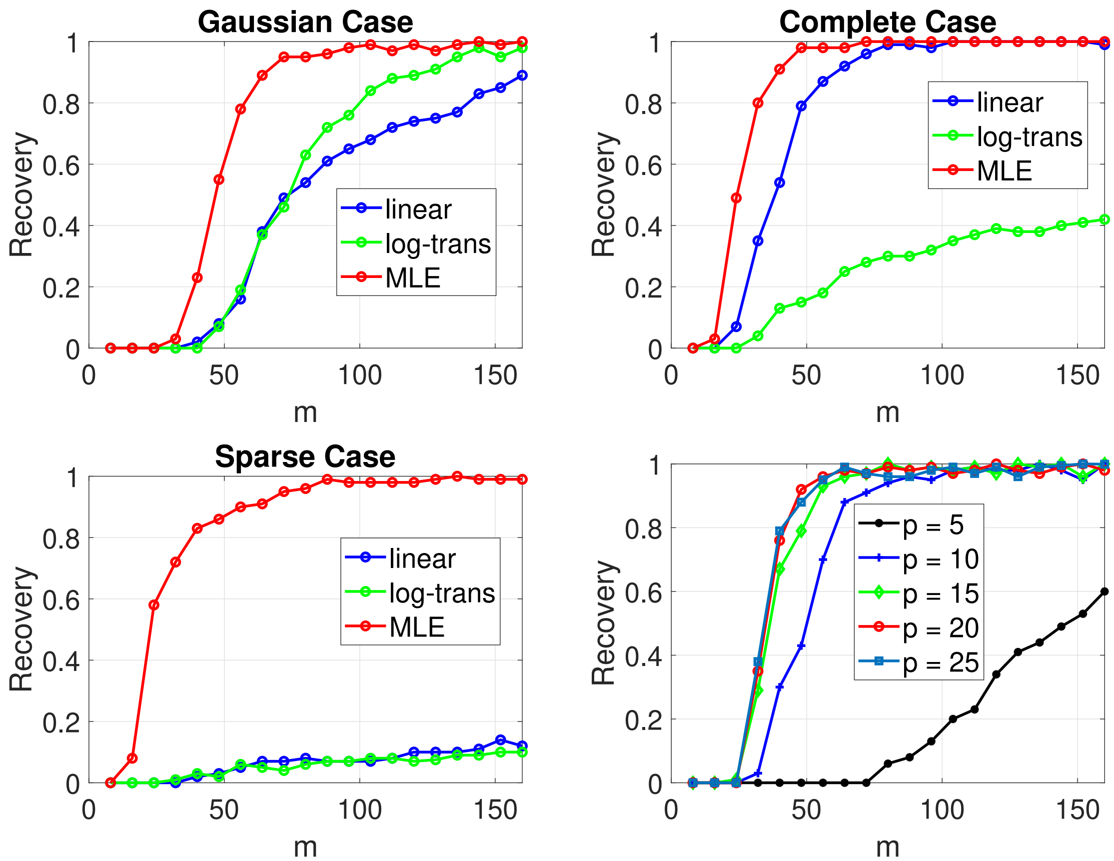

Figure 3. The top left, top right and bottom left plots show the curves of label permutation recovery under different design settings by using different estimation methods. The bottom right plot shows the permutation recovery curves under different p.

C. Real Data Example

In this section we apply our methods to a real financial dataset. The Dow Jones Industrial Average is a stock market index that measures the stock performance of 30 large companies listed on stock exchanges in the United States. The dataset consists of weekly price for each of thirty stocks in the first half of year 2011. There are 13 data columns in total, including open price, close price, volume, percent of change in price, percent of return of next dividend, etc.. Table 4 shows the data of first 5 weeks in 2011 for American Airline (AA). We pre-processed the dataset to transform it into a suitable form. We set $n = 30$ and $m = 25$ since there are 30 different stocks and 25 different dates. For each $i \in \{1, \ldots, 30\}$, we let y_{il} be the close price of ith on lth date (round to integer). We construct the design matrix by letting $p = 4$. We set X_{i1} to be the log of average open price for ith stock, set X_{i2} to be the log of average volume for ith stock, set X_{i3} to be the the percent of return of next dividend and set $X_{i4} = 1$ to incorporate the intercept term. We further scale columns $X[:, 1]$, $X[:, 2]$ and $X[:, 3]$ to make them have mean 0 and standard deviation 1.

We set the different values of observation rates (i.e. $q = 0.4 - 1.0$) and make different proportions of stocks assigned with wrong labels (i.e $h = d(\mathbf{I}, \Pi^\natural) = 0, 2, 4, 8, 12$). We fit the data by using Algorithm 4 and Algorithm 5 respectively. The results are given in Table 5 and Table 6. As we can see, both maximum likelihood estimation algorithm and two-step algorithm work well when the number of wrong labels is small and observation rate is 1. When the number of wrong labels gets larger, the maximum likelihood estimation algorithm is more robust, while two-step algorithm has a vanishing chance to recover the label permutation perfectly. On the other hand, when the observation rate decreases, the maximum likelihood estimation algorithm becomes less competitive.

Table 4. Illustration of Dow Jones Industrial Average dataset.

	stock	date	open	close	volume	…	percent return next dividend
1	AA	1/7/2011	\$15.82	\$16.42	239655616	…	0.182704
2	AA	1/14/2011	\$16.71	\$15.97	242963398	…	0.187852
3	AA	1/21/2011	\$16.19	\$15.79	138428495	…	0.189994
4	AA	1/28/2011	\$15.87	\$16.13	151379173	…	0.185989
5	AA	2/4/2011	\$16.18	\$17.14	154387761	…	0.175029
⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮

Table 5. Dow Jones Index Data: Permutation Recovery by Maximum Likelihood Estimation

	$q=0.4$	$q=0.5$	$q=0.6$	$q=0.7$	$q=0.8$	$q=0.9$	$q=1.0$
$h=0$	0.06	0.24	0.39	0.43	0.52	0.50	1.00
$h=2$	0.08	0.19	0.33	0.43	0.54	0.61	0.80
$h=4$	0.03	0.17	0.26	0.38	0.52	0.50	0.55
$h=8$	0.01	0.07	0.17	0.21	0.31	0.36	0.35
$h=12$	0.01	0.04	0.09	0.12	0.16	0.16	0.16

Table 6. Dow Jones Index Data: Permutation Recovery by "Two-step" Algorithm

	$q=0.4$	$q=0.5$	$q=0.6$	$q=0.7$	$q=0.8$	$q=0.9$	$q=1.0$
$h=0$	0.83	0.90	0.94	0.99	0.98	0.99	1.00
$h=2$	0.54	0.57	0.66	0.67	0.70	0.73	0.72
$h=4$	0.12	0.13	0.19	0.22	0.26	0.26	0.29
$h=8$	0.00	0.00	0.00	0.00	0.01	0.01	0.03
$h=12$	0.00	0.00	0.00	0.00	0.00	0.00	0.01

D. Explanation of Four Examples

In this appendix, we provide detailed explanations for examples given in Section 3.

Example 3.2. In this case, by convexity of ψ, we find that

$$\Delta_{ij} = \langle \psi'(\boldsymbol{\lambda}_i) \circ \boldsymbol{\lambda}_i - \psi(\boldsymbol{\lambda}_i)\rangle - \langle \psi'(\boldsymbol{\lambda}_i) \circ \boldsymbol{\lambda}_j - \psi(\boldsymbol{\lambda}_j)\rangle \geq \frac{1}{2}\kappa_0 \sum_l (x[i]B^\sharp[l] - x[j]B^\sharp[l]|)^2 \geq \frac{m\kappa_0}{2} x_{gap,ij}^2 b_1^2,$$

where κ_0 is the minimum $\psi''(x)$ over the range of $x = z_1 \cdot z_2$ with $z_1 \in [a_1, a_2]$ and $z_2 \in [b_1, b_2]$. We also have

$$v_{ij} = \sum_{l=1}^m \psi''(\boldsymbol{\lambda}_i[l])(\boldsymbol{\lambda}_j[l] - \boldsymbol{\lambda}_i[l])^2 \leq m\kappa_1 b_2^2 x_{gap,ij}^2,$$

where κ_0 is the maximum $\psi''(x)$ over the range of $x = z_1 z_2$ with $z_1 \in [a_1, a_2]$ and $z_2 \in [b_1, b_2]$.

Therefore, even with the knowledge of true parameter matrix $B^\sharp$, we still need large m to ensure the recovery of $\Pi^\sharp$. That is,

$$\frac{m\kappa_0}{2} x_{gap,ij}^2 b_1^2 \gtrsim \sqrt{(\log n) m\kappa_1 b_2^2 x_{gap,ij}^2}$$

for any pair of i, j. By simplification, we require $m \geq \max_{i,j} K \frac{\log n}{x_{gap,ij}^2} \approx n^4 \log n$ with some constant K.

Example 3.3. In this case, we define $w_{l,ij} := \mathbf{x}_i^T \mathbf{b}_l - \mathbf{x}_j^T \mathbf{b}_l$ and $d_{ij} := \sum_k \mathbf{1}\{X[i,k] \neq X[j,k]\}$ ($d_{ij} \geq 1$). We can find that $w_{l,ij}^2 / d_{ij}$ has mean 1 and variance $O(1)$. Then we have $\sum_{l=1}^m w_{l,ij}^2 = \Theta_p(m d_{ij})$.

Note that $\psi(x)$ is strictly convex, then there is a constant κ_0 such that $\psi(y) \geq \psi(x) + \psi'(x)(y-x) + \frac{\kappa_0}{2}(y-x)^2$ for any x, y.

By the same reason in Example 3.2, for any $i \neq j$, we have

$$\Delta_{ij} = \langle \psi'(\boldsymbol{\lambda}_i) \circ \boldsymbol{\lambda}_i - \psi(\boldsymbol{\lambda}_i) \rangle - \langle \psi'(\boldsymbol{\lambda}_i) \circ \boldsymbol{\lambda}_j - \psi(\boldsymbol{\lambda}_j) \rangle \geq \frac{\kappa_0}{2} \sum_{l=1}^{m} w_{l,ij}^2$$

held for some constant c and also have

$$v_{ij} = \sum_{l=1}^{m} \psi''(\boldsymbol{\lambda}_i[l])(\boldsymbol{\lambda}_j[l] - \boldsymbol{\lambda}_i[l])^2 \leq \kappa_1 \sum_{l=1}^{m} w_{l,ij}^2.$$

Therefore, when the true parameter matrix $B^\sharp$ is known, we need large m to ensure the recovery of $\Pi^\sharp$. That is,

$$\frac{\kappa_0}{2} \sum_{l=1}^{m} w_{l,ij}^2 \gtrsim \sqrt{(\log n)\kappa_1 \sum_{l=1}^{m} w_{l,ij}^2}.$$

By simplification, we require

$$\sum_{l=1}^{m} w_{l,ij}^2 \geq \frac{2\kappa_1}{\kappa_0} \log n.$$

Using $\sum_{l=1}^{m} w_{l,ij}^2 = \Theta_p(m d_{ij})$, it suffices to require $m \gtrsim \max_{ij} \frac{\log n}{d_{ij}} = O(\log n)$.

Example 3.4. In this case, by repeating the same procedure as in Example 3.3, we require

$$\sum_{l=1}^{m} w_{l,ij}^2 \geq \frac{2\kappa_1}{\kappa_0} \log n.$$

It suffices to find the lower bound of $\sum_{l=1}^{m} w_{l,ij}^2$. In this case, with high probability, $w_{l,ij}^2$ is lower bounded by $ca_1^2 d_{ij}$, where $d_{ij} = \sum_l \mathbf{1}\{X[i,l] - X[j,l] \neq 0\} \geq 1$ according to the assumption that each row of X has different support. Thus $\sum_{l=1}^{m} w_{l,ij}^2$ is bounded below by $ca_1^2 m$. We hence require that $m \gtrsim \log n$ for perfect recovery.

Example 3.5. Following the same reason in Example 3.3, we require

$$\sum_{l=1}^{m} w_{l,ij}^2 \geq \frac{2\kappa_1}{\kappa_0} \log n.$$

It also suffices to find the lower bound of $\sum_{l=1}^{m} w_{l,ij}^2$. In this case, given fixed $\mathbf{x}_i, \mathbf{x}_j$, $w_{l,ij}^2$, $w_{l,ij}^2 / \|\mathbf{x}_i - \mathbf{x}_j\|^2$ follows a Chi-square distribution with degree 1 bounded by $ca_1^2 d_{ij}$. With high probability, it holds $\sum_l w_{l,ij}^2 = \Theta_p(m\|\mathbf{x}_i - \mathbf{x}_j\|^2)$ Therefore, it suffices to have $m \gtrsim \max_{i,j} \frac{\log n}{\|\mathbf{x}_i - \mathbf{x}_j\|^2} = \Theta_p(\log n)$.

E. Explanation of $\Delta(X, B^\sharp, \Pi^\sharp, \Pi)$

By the definition, we can see that there is no explicit form for $\Delta(X, B^\sharp, \Pi^\sharp, \Pi)$. In this appendix, we provide a discussion on the lower bound of $\Delta(X, B^\sharp, \Pi^\sharp, \Pi)$.

Note that we can always rewrite $\Pi^\sharp X$ as X and treat $\Pi(\Pi^\sharp)^{-1}$ as Π. We then assume that $\Pi^\sharp = \mathbf{I}$ for the sake of simplicity. Moreover, we only need to consider $m = 1$ by noticing that $\Lambda(\Pi, B)$ can be written as the separate function of each column of B.

Take any $\Pi \neq \mathbf{I}$ and let $h := d(\Pi, \mathbf{I})$. Here, without loss of generality, we assume that $\Pi(i) = i$ for any $i > h$. By the definition that $\Lambda(\Pi) = \max_{\mathbf{b}} \Lambda(\Pi, \mathbf{b}) = \max_{\mathbf{b}} \sum_{i=1}^n \Lambda_i(\Pi, \mathbf{b})$, and $\mathbf{b}^\sharp$ is the true parameter, we then have that

$$\begin{aligned}
\Lambda(\mathbf{I}) - \Lambda(\Pi) &= \Lambda(\mathbf{I}, \mathbf{b}^\sharp) - \max_{\mathbf{b}} \Lambda(\Pi, \mathbf{b}) \\
&= \sum_{i=1}^n \Lambda_i(\mathbf{I}, \mathbf{b}^\sharp) - \max_{\mathbf{b}} \sum_{i=1}^n \Lambda_i(\Pi, \mathbf{b}) \\
&\geq \sum_{i=1}^n \Lambda_i(\mathbf{I}, \mathbf{b}^\sharp) - (\max_{\mathbf{b}} \sum_{i=1}^h \Lambda_i(\Pi, \mathbf{b}) + \max_{\mathbf{b}} \sum_{i=h+1}^n \Lambda_i(\Pi, \mathbf{b})) \\
&\geq \sum_{i=1}^h \Lambda_i(\mathbf{I}, \mathbf{b}^\sharp) - \max_{\mathbf{b}} \sum_{i=1}^h \Lambda_i(\Pi, \mathbf{b}) \\
&\geq \min_{\mathbf{b}} \sum_{i=1}^h \{\Lambda_i(\mathbf{I}, \mathbf{b}^\sharp) - \Lambda_i(\Pi, \mathbf{b})\} \\
&\geq \lambda_0 \min_{\mathbf{b}} \sum_{i=1}^h (X[i, :]\mathbf{b}^\sharp - X[\Pi(i),]\mathbf{b})^2 \\
&\geq \lambda_0 \min_{\mathbf{b}} d_{gap}^2, \qquad (43)
\end{aligned}$$

where $d_{gap} := \min_{\mathbf{b}} \|X_{\Pi,1}\mathbf{b} - X_1\mathbf{b}^\sharp\|$ with $X_1 := X[1 : h,]$ and $X_{\Pi,1} = (\Pi X)[1 : h,]$ and λ_0 is equal to $\min_i\{\psi''(\mathbf{x}_i^T\mathbf{b}^\sharp), \psi''(\mathbf{x}_i^T\mathbf{b}(\Pi))\}$. Moreover, d_{gap} admits an explicit form, which is,

$$d_{gap} = \|\mathbf{P}_{X_{\Pi,1}} X_1 \mathbf{b}^\sharp\|,$$

where $\mathbf{P}_{X_{\Pi,1}} = I - X_{\Pi,1}(X_{\Pi,1}^T X_{\Pi,1})^{-1} X_{\Pi,1}^T$.

When $p = 1$, we know that $X_{\Pi,1}^T X_{\Pi,1}$ is equal to $\|X_{\Pi,1}\|^2$. Thus

$$\begin{aligned}
\|(X_{\Pi,1}^T X_{\Pi,1})^{-1} X_{\Pi,1}^T X_1\| &\leq \|X_{\Pi,1}^T X_1\| / \|X_{\Pi,1}\|^2 = 1 - \frac{\|X_{\Pi,1}\|^2 - \|X_{\Pi,1}^T X_1\|}{\|X_{\Pi,1}\|^2} \\
&= (1 - \frac{\|X_{\Pi,1}\|^2 - \|X_{\Pi,1}^T X_1\|}{\|X_{\Pi,1}\|^2}) \\
&= (1 - \frac{\|X_{\Pi,1}\|^2 - \|X_{\Pi,1}^T X_1\| + \|X_1\|^2 - \|X_1^T X_{\Pi,1}\|}{2\|X_{\Pi,1}\|^2}) \\
&\leq (1 - \frac{\|X_{\Pi,1} - X_1\|^2}{2\|X_{\Pi,1}\|^2}).
\end{aligned}$$

Thus

$$d_{gap} = \|X_1\mathbf{b}^\sharp - X_{\Pi,1}(X_{\Pi,1}^T X_{\Pi,1})^{-1} X_{\Pi,1}^T X_1 \mathbf{b}^\sharp\| \geq \underbrace{\|X_1\mathbf{b}^\sharp\|}_{\Omega(\sqrt{h})} \underbrace{\frac{\|X_{\Pi,1} - X_1\|^2}{2\|X_{\Pi,1}\|^2}}_{\Theta(1)} = \Omega(\sqrt{h}).$$

Therefore $\Lambda(\mathbf{I}) - \Lambda(\Pi) \geq c_0 m d(\mathbf{I}, \Pi)$ for some constant c_0 which is related to the design matrix X.

When $p > 1$, there exists a rotation matrix W such that $W\mathbf{b}^\sharp = \mathbf{e}_1\|\mathbf{b}^\sharp\|$ ($\mathbf{e}_1$ is a vector with all entries being zero but the first entry being 1). Write $X_1 = \tilde{X}_1 W$. Then,

$$d_{gap} = \min_{\mathbf{b}} \|X_{\Pi,1}\mathbf{b} - X_1\mathbf{b}^\sharp\| = \min_{\mathbf{b}} \|\tilde{X}_{\Pi,1}\mathbf{b} - \tilde{X}_1\mathbf{e}_1\|\mathbf{b}^\sharp\|\| = \|\mathbf{b}^\sharp\| \min_{\mathbf{b}} \|\tilde{X}_{\Pi,1}\mathbf{b} - \tilde{X}_1\mathbf{e}_1\|.$$

Thus, we have $d_{gap} = \|\mathbf{b}^\sharp\|\|(I - \tilde{X}_{\Pi,1}(\tilde{X}_{\Pi,1}^T\tilde{X}_{\Pi,1})^{-1}\tilde{X}_{\Pi,1}^T)\tilde{X}_1\mathbf{e}_1\| \geq c_0\|\mathbf{b}^\sharp\|\|\tilde{X}_1 e_1\| = c_0\Omega(\sqrt{h})$, where c_0 is the distance from $\tilde{X}_1 e_1/\|\tilde{X}_1 e_1\|$ to the space spanned by $\tilde{X}_{\Pi,1}$. Thus $\Lambda(\mathbf{I}) - \Lambda(\Pi) \geq c_0' m d(\mathbf{I}, \Pi)$ by adjusting the constant c_0'.

F. Proof of Results when B is Known: Theorem 3.1

In this section, we prove the results when B is known. We additionally use $A[i,:]/A[:,j]$ to represent the ith row/jth column of matrix A; $\text{diag}(\mathbf{a})$ is the diagonal matrix with lth diagonal element being $\mathbf{a}[l]$; $\|A\|_F$ is the Frobenius norm of matrix A; $\|A\|_{col} := \max_j \|A[:,j]\|$; $\sigma_{min}(A)/\sigma_{max}(A)$ represents the minimum/maximum positive singular value of matrix A. Note that the log-likelihood function is a *separable* function of each column of parameter matrix B, therefore we sometimes also *treat B as a column vector* for notational simplicity.

To prove the result, we only need to show the following probability,

$$P(\sup_{\Pi \neq \Pi^\sharp} L(\Pi, B) \geq L(\Pi^\sharp, B)), \tag{44}$$

goes to zero as n and m increase. The naive union bound will give an upper bound,

$$\sum_{\Pi \neq \Pi^\sharp} P(L(\Pi, B) \geq L(\Pi^\sharp, B)).$$

Note that there are $n!$ possible permutations. The above probability could be exponentially large.

Fortunately, we can find that log-likelihood $L(\Pi, B) = \sum_{i=1}^n \langle -\psi((\Pi X B)[i,:]) + Y \circ (\Pi X B)[i,:] \rangle$ is an additive function of X's rows. Therefore, (44) is bounded by

$$\begin{aligned} &\leq P(\max_{j \neq i} \langle -\psi((XB)[j,:]) + (Y[i,:] \circ (XB)[j,:]) \rangle \geq \langle -\psi((XB)[i,:]) + (Y \circ XB)[i,:] \rangle) \\ &\leq \sum_{j \neq i} P(\langle -\psi((XB)[j,:]) + (Y[i,:] \circ (XB)[j,:]) \rangle \geq \langle -\psi((XB)[i,:]) + (Y \circ XB)[i,:] \rangle) \end{aligned}$$

Next we bound each term, $P(\langle -\psi((XB)[j,:]) + (Y[i,:] \circ (XB)[j,:]) \rangle \geq \langle -\psi((XB)[i,:]) + (Y \circ XB)[i,:] \rangle)$, in above inequality.

Recall the definition of $\boldsymbol{\lambda}_i$, we thus have

$$\begin{aligned} & \mathbb{E}\langle \mathbf{y}_i \circ \boldsymbol{\lambda}_i - \psi(\boldsymbol{\lambda}_i) \rangle \\ = & \langle \psi'(\boldsymbol{\lambda}_i) \circ \boldsymbol{\lambda}_i - \psi(\boldsymbol{\lambda}_i) \rangle. \end{aligned} \tag{45}$$

It can be checked that

$$\Delta_{ij} = \langle \psi'(\boldsymbol{\lambda}_i) \circ \boldsymbol{\lambda}_i - \psi(\boldsymbol{\lambda}_i) \rangle - \langle \psi'(\boldsymbol{\lambda}_i) \circ \boldsymbol{\lambda}_j - \psi(\boldsymbol{\lambda}_j) \rangle \geq 0 \tag{46}$$

for any convex function ψ.

For any $i \neq j$, we next calculate the variance of $\langle \mathbf{y}_i \circ \boldsymbol{\lambda}_i - \psi(\boldsymbol{\lambda}_i) \rangle - \langle \mathbf{y}_i \circ \boldsymbol{\lambda}_j - \psi(\boldsymbol{\lambda}_j) \rangle$.

$$\begin{aligned} & \text{var}(\langle \mathbf{y}_i \circ \boldsymbol{\lambda}_i - \psi(\boldsymbol{\lambda}_i) \rangle - \langle \mathbf{y}_i \circ \boldsymbol{\lambda}_j - \psi(\boldsymbol{\lambda}_j) \rangle) \\ \leq & \sum_{l=1}^m \text{var}(Y[i,l](\boldsymbol{\lambda}_i[l] - \boldsymbol{\lambda}_j[l])) \\ = & \sum_{l=1}^m \psi''(\boldsymbol{\lambda}_i[l])(\boldsymbol{\lambda}_i[l] - \boldsymbol{\lambda}_j[l])^2 =: v_{ij}. \end{aligned} \tag{47}$$

To characterize the difference between Δ_{ij} and $\langle Y[i,:] \circ \boldsymbol{\lambda}_i - \psi(\boldsymbol{\lambda}_i) \rangle - \langle Y[i,:] \circ \boldsymbol{\lambda}_j - \psi(\boldsymbol{\lambda}_j) \rangle$, we use the following lemma.

Lemma F.1. *There exists a constant c_ψ (may depends on ψ) such that,*

$$P(|\langle \mathbf{y}_i \circ \boldsymbol{\lambda}_i - \psi(\boldsymbol{\lambda}_i) \rangle - \langle \mathbf{y}_i \circ \boldsymbol{\lambda}_j - \psi(\boldsymbol{\lambda}_j) \rangle - \Delta_{ij}| \geq v_{ij} x) \leq \exp\{-v_{ij}(\min\{x, c_\psi\})^2/4\}, \tag{48}$$

holds for any $x > 0$.

Proof of Lemma F.1. With some calculations, we know

$$\begin{aligned}
& P(|\langle \mathbf{y}_i \circ \boldsymbol{\lambda}_i - \psi(\boldsymbol{\lambda}_i)\rangle - \langle \mathbf{y}_i \circ \boldsymbol{\lambda}_j - \psi(\boldsymbol{\lambda}_j)\rangle - \Delta_{ij}| \geq v_{ij}x) \\
= \; & P(|\langle (\mathbf{y}_i - \psi'(\boldsymbol{\lambda}_i)) \circ (\boldsymbol{\lambda}_i - \boldsymbol{\lambda}_j)\rangle| \geq v_{ij}x) \\
\leq \; & \inf_{0<t<c'_\psi} \exp\{-tv_{ij}x\}\exp\{v_{ij}t^2\} \;\; \text{(using MGF and Markov inequality)} \quad (49) \\
\leq \; & \exp\{-v_{ij}x^2/4\} \;\; (\text{taking } t = \frac{x}{2}), \quad (50)
\end{aligned}$$

where (49) utilizes the property of moment generating function of generalized linear model. That is, it is well known that $\mathbb{E}[\exp\{tY\}] = \exp\{\psi(\lambda + t) - \psi(\lambda)\}$, where the density of random variable Y is proportional to $\exp\{y\lambda - \psi(\lambda)\}$.

By calculations, we know

$$\begin{aligned}
& \exp\{t\langle (\mathbf{y}_i - \psi'(\boldsymbol{\lambda}_i)) \circ (\boldsymbol{\lambda}_i - \boldsymbol{\lambda}_j)\rangle\} \\
= \; & \prod_{l=1}^m \exp\{\psi(\boldsymbol{\lambda}_i[l] + t(\boldsymbol{\lambda}_i[l] - \boldsymbol{\lambda}_j[l])) - \psi(\boldsymbol{\lambda}_i[l]) - t\psi'(\boldsymbol{\lambda}_i[l])(\boldsymbol{\lambda}_i[l] - \boldsymbol{\lambda}_j[l])\} \\
\leq \; & \prod_{l=1}^m \exp\{\psi''(\boldsymbol{\lambda}_i[l])(\boldsymbol{\lambda}_i[l] - \boldsymbol{\lambda}_j[l])^2 t^2\} \;\; (\forall t \text{ satisfying } \frac{1}{2}\psi''(\boldsymbol{\lambda}_i[l]) > |\psi'''(\boldsymbol{\lambda}_i[l])(\boldsymbol{\lambda}_i[l] - \boldsymbol{\lambda}_j[l])t|) \\
= \; & \exp\{t^2 \sum_{l=1}^m \psi''(\boldsymbol{\lambda}_i[l])(\boldsymbol{\lambda}_i[l] - \boldsymbol{\lambda}_j[l])^2\} \\
= \; & \exp\{v_{ij}t^2\}.
\end{aligned}$$

In other words,

$$\exp\{t\langle (\mathbf{y}_i - \psi'(\boldsymbol{\lambda}_i)) \circ (\boldsymbol{\lambda}_i - \boldsymbol{\lambda}_j)\rangle\} \leq \exp\{v_{ij}t^2\}$$

holds for any $0 < t < \min_{i,j,l} \frac{\frac{1}{2}\psi''(\boldsymbol{\lambda}_i[l])}{\psi'''(\boldsymbol{\lambda}_i[l])|\boldsymbol{\lambda}_i[l] - \boldsymbol{\lambda}_j[l]|} =: c'_\psi$. By taking $c_\psi = 2c'_\psi$, it completes the proof for any $0 < x < c_\psi$. Finally, by noticing that the left hand side of (55) is decreasing function of x ($x > 0$). We have

$$P(|\langle \mathbf{y}_i \circ \boldsymbol{\lambda}_i - \psi(\boldsymbol{\lambda}_i)\rangle - \langle \mathbf{y}_i \circ \boldsymbol{\lambda}_j - \psi(\boldsymbol{\lambda}_j)\rangle - \Delta_{ij}| \geq v_{ij}x) \leq \exp\{-v_{ij}c_\psi^2/4\}$$

for any $x > c_\psi$. This concludes the proof. □

By taking $x = \Delta_{ij}/v_{ij}$ in (55), this gives

$$\begin{aligned}
& P(\langle -\psi((XB)[j,:]) + (Y[i,:] \circ (XB)[j,:])\rangle \geq \langle -\psi((XB)[i,:]) + (Y \circ XB)[i,:]\rangle) \\
\leq \; & P(|\langle -\psi((XB)[j,:]) + (Y[i,:] \circ (XB)[j,:])\rangle - \langle -\psi((XB)[i,:]) + (Y \circ XB)[i,:]\rangle - \Delta_{ij}| \geq \Delta_{ij}) \\
\leq \; & \max\{\exp\{-\Delta_{ij}^2/(4v_{ij})\}, \exp\{-v_{ij}c_\psi^2/4\}\}. \quad (51)
\end{aligned}$$

Finally, by union bound with summing over all possible pairs of i and j, this completes the proof of Theorem 3.1.

G. Proof of Results with Knowledge that $d(\mathbf{I}, \Pi^\natural)$ is Small: Theorem 4.4

In this section, we prove the results when we have the prior knowledge that $d(\mathbf{I}, \Pi^\natural)$ is small. For ease of presentation, we treat $\Pi^\natural$ as $\mathbf{I}$. (By doing this, it will not change the technical difficulty.) In order to prove the recovery consistency, we need to control the quantity $\sup_{B \in B_\delta(B^\natural)} \Big\{(L(\mathbf{I}, B) - L(\Pi, B)) - (\Lambda(\mathbf{I}, B) - \Lambda(\Pi, B))\Big\}$ (where $B_\delta(B^\natural) := \{B : \|\mathbf{b}_l - \mathbf{b}_l^\natural\|_2 \leq \delta,\ l \in [m]\}$) and identify a δ, which is an upper bound of $\|B - B^\natural\|_{col}$ (recalling the definition of norm $\|A\|_{col} := \max_j \|A[:,j]\|$.)

We first to derive the high probability bound of $\sup_{B:\|B - B^\natural\|_{col} \leq \delta} \Big\{(L(\mathbf{I}, B) - L(\Pi, B)) - (\Lambda(\mathbf{I}, B) - \Lambda(\Pi, B))\Big\}$ through the following three lemma. In the following, constants c_ψ, c'_ψ, C, C' may vary from place to place.

Lemma G.1. *Define*

$$\begin{aligned} \textit{DIFF}(\mathbf{I}, \Pi, B) &:= (L(\mathbf{I}, B) - L(\Pi, B)) - (\Lambda(\mathbf{I}, B) - \Lambda(\Pi, B)) \\ &= \sum_l \sum_{i:\Pi(i)\neq i} \Big\{ Y[i,l](\mathbf{x}_i^T \mathbf{b}_l) - \psi(\mathbf{x}_i^T \mathbf{b}_l) - Y[i,l](\mathbf{x}_{\Pi(i)}^T \mathbf{b}_l) - \psi(\mathbf{x}_{\Pi(i)}^T \mathbf{b}_l) \Big\} \\ &\quad - \sum_l \sum_{i:\Pi(i)\neq i} \Big\{ -\psi(\mathbf{x}_i^T \mathbf{b}_l) + \psi'(\mathbf{x}_i^T \mathbf{b}_l)\mathbf{x}_i^T \mathbf{b}_l - (-\psi(\mathbf{x}_{\Pi(i)}^T \mathbf{b}_l) + \psi'(\mathbf{x}_{\Pi(i)}^T \mathbf{b}_l)\mathbf{x}_{\Pi(i)}^T \mathbf{b}_l) \Big\}. \end{aligned} \tag{52}$$

It holds

$$\begin{aligned} & |\textit{DIFF}(\mathbf{I}, \Pi, B) - \textit{DIFF}(\mathbf{I}, \Pi, B')| \\ \leq & \sum_l \sum_{i:\Pi(i)\neq i} \Big\{ 2x_{max} Y_{max} \delta + 2c_\psi x_{max} \delta \Big\} \leq c'_\psi m d(\mathbf{I}, \Pi) x_{max} \delta \end{aligned} \tag{53}$$

for some constants c_ψ, c'_ψ which depend on ψ and any B, B' satisfying $\|B - B'\|_{col} \leq \delta$.

Proof of Lemma G.1. By (52), we know

$$\begin{aligned} & |\text{DIFF}(\mathbf{I}, \Pi, B) - \text{DIFF}(\mathbf{I}, \Pi, B')| \\ = & \underbrace{\sum_l \sum_{i:\Pi(i)\neq i} \Big\{ Y[i,l](\mathbf{x}_i^T (\mathbf{b}_l - \mathbf{b}'_l)) - Y[i,l](\mathbf{x}_{\Pi(i)}^T (\mathbf{b}_l - \mathbf{b}'_l)) \Big\}}_{D1} \\ & \underbrace{- \sum_l \sum_{i:\Pi(i)\neq i} \Big\{ \psi'(\mathbf{x}_i^T \mathbf{b}_l)\mathbf{x}_i^T \mathbf{b}_l - \psi'(\mathbf{x}_i^T \mathbf{b}'_l)\mathbf{x}_i^T \mathbf{b}'_l - \psi'(\mathbf{x}_{\Pi(i)}^T \mathbf{b}_l)\mathbf{x}_{\Pi(i)}^T \mathbf{b}_l + \psi'(\mathbf{x}_{\Pi(i)}^T \mathbf{b}'_l)\mathbf{x}_{\Pi(i)}^T \mathbf{b}'_l \Big\}}_{D2}. \end{aligned} \tag{54}$$

It is easy to get that $D_1 \leq \sum_l \sum_{i:\Pi(i)\neq i} 2x_{max} Y_{max} \delta$ by Cauchy-Schwartz inequality. For D_2, we first observe that $|\psi'(x_1)x_1 - \psi'(x_2)x_2| \leq \sup_{x' \in [x_1, x_2]} \{\psi''(x')x' + \psi'(x')\} |x_1 - x_2|$ for any $x_1 < x_2 \in \mathbb{R}$. Moreover, we define

$$c_\psi := \max_{i,l} \sup_{x' \in [\mathbf{x}_i^T \mathbf{b}_l - x_{max}\delta, \mathbf{x}_i^T \mathbf{b}_l + x_{max}\delta]} \{\psi''(x')x' + \psi'(x')\}.$$

Therefore,

$$D_2 \leq \sum_l \sum_{i:\Pi(i)\neq i} 2c_\psi x_{max} \delta.$$

Combining the upper bounds of D_1 and D_2, we get the desire result by adjusting constant c'_ψ. □

Lemma G.2. *There exists a constant c_ψ (may depends on ψ) such that,*

$$P(|(L(\mathbf{I}, B^\sharp) - L(\Pi, B^\sharp)) - (\Lambda(\mathbf{I}, B^\sharp) - \Lambda(\Pi, B^\sharp))| \geq v_{\Pi,partial} x) \leq \exp\{-v_{\Pi,partial} (\min\{x, c_\psi\})^2 / 4\}, \tag{55}$$

holds for any $x > 0$

Proof of Lemma G.2. Proof is the same as that of Lemma F.1 by treating $v_{\Pi,partial}$ as v_{ij} there. □

Lemma G.3. *For any x and δ such that $c'_\psi m d(\mathbf{I}, \Pi) x_{max} \delta < v_{\Pi,partial} x$ and $x < c_\psi$, it holds*

$$\begin{aligned} & P(\sup_{B \in B_\delta(B^\sharp)} |(L(\mathbf{I}, B) - L(\Pi, B)) - (\Lambda(\mathbf{I}, B) - \Lambda(\Pi, B))| \geq 2v_{\Pi,partial} x) \\ \leq & \exp\{-\frac{1}{4} v_{\Pi,partial} x^2\}. \end{aligned} \tag{56}$$

Here constant c_ψ is the same as that in Lemma G.2 and c'_ψ is the same as that in Lemma G.1.

Proof of Lemma G.3.

$$\begin{aligned}
& P(\sup_{B\in B_\delta(B^\sharp)} |(L(\mathbf{I},B)-L(\Pi,B))-(\Lambda(\mathbf{I},B)-\Lambda(\Pi,B))| \geq 2v_{\Pi,partial}x) \\
\leq\ & P(|(L(\mathbf{I},B^\sharp)-L(\Pi,B^\sharp))-(\Lambda(\mathbf{I},B^\sharp)-\Lambda(\Pi,B^\sharp))| \geq v_{\Pi,partial}x) \\
& +P(|\text{DIFF}(\mathbf{I},\Pi,B)-\text{DIFF}(\mathbf{I},\Pi,B')| \geq v_{\Pi,partial}x) \\
\leq\ & \exp\{-\frac{1}{4}v_{\Pi,partial}x^2\}. \text{ (Use Lemma G.2 and the following fact.)} \qquad (57)
\end{aligned}$$

The last inequality holds due to the fact that

$$|\text{DIFF}(\mathbf{I},\Pi,B)-\text{DIFF}(\mathbf{I},\Pi,B')| \leq c'_\psi m d(\mathbf{I},\Pi)x_{max}\delta < v_{\Pi,partial}x$$

leading to

$$P(|\text{DIFF}(\mathbf{I},\Pi,B)-\text{DIFF}(\mathbf{I},\Pi,B')| \geq v_{\Pi,partial}x) = 0.$$

□

Main Idea of Proof: Suppose we have already known that the estimator $\hat{B}$ which is close to the truth $B^\sharp$, i.e., in the δ-neighborhood of $B^\sharp$. Then by Lemma G.3, we know that

$$P(\sup_{B\in B_\delta(B^\sharp)} |(L(\mathbf{I},B)-L(\Pi,B))-(\Lambda(\mathbf{I},B)-\Lambda(\Pi,B))| \geq 2v_{\Pi,partial}x) \qquad (58)$$

vanishes for any fixed Π, where $\delta(=\delta^*)$ is a sufficiently small constant which is determined later.

Hence, with straight forward calculations, we have

$$\begin{aligned}
& P(\max_{\Pi}\sup_{B\in B_\delta(B^\sharp)} |(L(\mathbf{I},B)-L(\Pi,B))-(\Lambda(\mathbf{I},B)-\Lambda(\Pi,B))| \geq 2v_{\Pi,partial}x) \\
\leq\ & \sum_{\Pi} P(|(L(\mathbf{I},B^\sharp)-L(\Pi,B^\sharp))-(\Lambda(\mathbf{I},B^\sharp)-\Lambda(\Pi,B^\sharp))| \geq v_{\Pi,partial}x) \text{ (union bound)} \\
\leq\ & \sum_h \sum_{\Pi:d(\Pi,\mathbf{I})=h} \exp\{-v_{\Pi,partial}x^2/4\} \text{ (Lemma G.2)} \\
\leq\ & \sum_h n!/(n-h)!\cdot\exp\{-v_{\Pi,partial}x^2/4\} \\
\leq\ & \sum_h n^h\cdot\exp\{-hv_{min}x^2/4\} \text{ (using fact: } |\{\Pi:d(\Pi,\mathbf{I})=h\}|\leq n!/(n-h)!\leq n^h) \\
=\ & \sum_h \exp\{-h(v_{min}x^2/4-\log n)\} \\
\leq\ & \frac{\exp\{-2(v_{min}x^2/4-\log n)\}}{1-\exp\{-(v_{min}x^2/4-\log n)\}}, \qquad (59)
\end{aligned}$$

where $v_{min} := \min_{i,j}\sum_l \boldsymbol{\lambda}^\sharp_{il}\log^2(\boldsymbol{\lambda}^\sharp_{jl}/\boldsymbol{\lambda}^\sharp_{il})$.

By (59), we know that

$$L(\Pi,\hat{B}) \leq L(\mathbf{I},\hat{B})-\Lambda(\mathbf{I},\hat{B})+\Lambda(\Pi,\hat{B})+x\cdot v_{\Pi,partial} \qquad (60)$$

via taking

$$x := C\sqrt{\log n/v_{min}}$$

(C is some constant) with probability going to 1 as $n\to\infty$. This tells us that if we can show that

$$\Lambda(\mathbf{I},\hat{B})-\Lambda(\Pi,\hat{B})-x\cdot v_{\Pi,partial} > 0 \qquad (61)$$

for any Π with $d(\mathbf{I},\Pi)\leq h_{max}$ and $\hat{B}$ within δ-neighborhood of $B^\sharp$. Together with (60), it implies

$$L(\Pi,\hat{B}) \leq L(\mathbf{I},\hat{B})$$

with probability going to 1. Then we can conclude that $\hat{\Pi}=\mathbf{I}$ which gives the desired result.

In the following subsections, we characterize the neighborhood radius δ in G.1-G.2 and prove (61).

G.1. First bound of $\|\hat{B} - B^\sharp\|_{col}$

Since the likelihood function can be written as the sum of separate functions of $\mathbf{b}_1, \ldots, \mathbf{b}_m$, we only need to focus on single $\mathbf{b}_l$ for $l \in [m]$ one by one. This reduces to the case: $m = 1$. Then the estimator $\hat{\mathbf{b}}$ is

$$\hat{\mathbf{b}} = \arg\max_{\mathbf{b}} \{\langle -\psi(X\mathbf{b}) + Y \circ X\mathbf{b} \rangle\}. \tag{62}$$

When $d(\mathbf{I}, \Pi^\sharp) \leq h_{max}$, we aim to show $\hat{\mathbf{b}}$ is a consistent estimator of $\mathbf{b}^\sharp$.

For simplicity, we can assume $\Pi^\sharp(i) = i$ for $i > h_{max}$ and let $L(\mathbf{b}) = \langle -\psi(X\mathbf{b}) + Y \circ X\mathbf{b} \rangle$. In the following, we aim to find a δ_n such that for any $\mathbf{b}$ with $\|\mathbf{b} - \mathbf{b}^\sharp\| \geq \delta_n$, it holds $L(\mathbf{b}) < L(\mathbf{b}^\sharp)$. By the definition that $L(\hat{\mathbf{b}}) \geq L(\mathbf{b}^\sharp)$, we will arrive at $\|\hat{\mathbf{b}} - \mathbf{b}^\sharp\| \leq \delta_n$.

Following curvature inequality technique (see (90)) described in Section H.1. We can specifically take

$$\delta_n^2 := C\gamma_{1p} \frac{\sqrt{v_{b^\sharp}} + \psi^\sharp_{max} h_{max}}{(n - h_{max})\psi^\sharp_{min}}$$

with some large constant C and $v_b := \sum_{i=1}^n \psi''(\lambda_i^\sharp)(\lambda_i(\mathbf{b}))^2$. With this choice of δ_n, we can easily check that

$$L(\mathbf{b}^\sharp) - L(\mathbf{b}) > 0$$

for any $\mathbf{b}$ with $\|\mathbf{b} - \mathbf{b}^\sharp\| = \delta_n$ when $p = O(n^a)$ ($a < \frac{1}{2}$) and $\delta_n = o_p(1)$. By the concavity of likelihood function, we then know that the estimator $\hat{\mathbf{b}}$ must lie in the ball $\{\mathbf{b} : \|\mathbf{b} - \mathbf{b}^\sharp\| \leq \delta_n\}$.

G.2. Second bound of $\|\hat{B} - B^\sharp\|_{col}$

The first bound of $\|\hat{B} - B^\sharp\|_{col}$ implies that $\|\hat{B} - B^\sharp\|_{col} = o_p(1)$. In this section, we can further improve this bound.

We consider the Taylor expansion of $L(\mathbf{b})$ at value $\mathbf{b}^\sharp$. Then, it can be computed that

$$\mathbf{0} = \nabla L(\hat{\mathbf{b}}) = \nabla L(\mathbf{b}^\sharp) + \nabla^2 L(\bar{\mathbf{b}})(\hat{\mathbf{b}} - \mathbf{b}^\sharp) \tag{63}$$

where $\bar{\mathbf{b}}$ is some point between $\hat{\mathbf{b}}$ and $\mathbf{b}^\sharp$. By the formula $\nabla^2 L(\mathbf{b}) = \sum_{i=1}^n \psi''(\mathbf{x}_i^T\mathbf{b})\mathbf{x}_i\mathbf{x}_i^T$, we then know that

$$\begin{aligned} |\nabla^2 L(\mathbf{b}^\sharp) - \nabla^2 L(\bar{\mathbf{b}})| &= |\sum_{i=1}^n \psi''(\mathbf{x}_i^T\mathbf{b}^\sharp)\mathbf{x}_i\mathbf{x}_i^T(1 - \psi''(\mathbf{x}_i^T\bar{\mathbf{b}})/\psi''(\mathbf{x}_i^T\mathbf{b}^\sharp))| \\ &\leq |\sum_{i=1}^n \psi''(\mathbf{x}_i^T\mathbf{b}^\sharp) \cdot o_p(1) \cdot \mathbf{x}_i\mathbf{x}_i^T|, \end{aligned} \tag{64}$$

since $\|\bar{\mathbf{b}} - \mathbf{b}^\sharp\| \leq \|\hat{\mathbf{b}} - \mathbf{b}^\sharp\| = o_p(1)$ according to the first bound result. Then we know that

$$\sigma_{min}(\nabla^2 L(\bar{\mathbf{b}})) = (1 + o_p(1))\sigma_{min}(\nabla^2 L(\mathbf{b}^\sharp)) \geq \sigma_{min}(\nabla^2 L(\mathbf{b}^\sharp))/2.$$

We thus have

$$\|\hat{\mathbf{b}} - \mathbf{b}^\sharp\| = \|(\nabla^2 L(\bar{\mathbf{b}}))^{-1}\nabla L(\mathbf{b}^\sharp)\| \leq \frac{2}{\sigma_{min}(\nabla^2 L(\mathbf{b}^\sharp))}\|\nabla L(\mathbf{b}^\sharp)\|. \tag{65}$$

For l-th ($l = 1, \ldots, p$) element of $\nabla L(\mathbf{b}^\sharp)$, we can find that

$$\nabla L(\mathbf{b}^\sharp)[l] = \sum_{i \leq h_{max}} \nabla L_i(\mathbf{b}^\sharp)[l] + \sum_{i > h_{max}} \nabla L_i(\mathbf{b}^\sharp)[l]. \tag{66}$$

The first term of (66) is bounded by

$$b_s := |\sum_{i \leq h_{max}} (Y[i] - \psi'(\mathbf{x}_i^T\mathbf{b}^\sharp))X[i, l]\}|,$$

which is order of $\log n \sum_{i \le h_{max}} (\psi''^{\sharp}_{max} \log n + \psi'^{\sharp}_{max})|X[i,l]|$. Here we use the observation that, for sub-exponential random variables,

$$Y[i,j] = O_p(\sqrt{\psi''^{\sharp}_{max}} \log n + \psi'^{\sharp}_{max}) = O_p(\psi''^{\sharp}_{max} \log n + \psi'^{\sharp}_{max}) \tag{67}$$

for all $i \in [n], j \in [m]$. In the following, without loss of generality, we treat all observed entries in Y are at most of order $\log n$.

The second term of (66) is bounded by $C\sqrt{\text{var}(\sum_{i>h_{max}} \nabla L_i(\mathbf{b}^\sharp)[l])}$ and the upper bound of $\text{var}(\sum_{i>h_{max}} \nabla L_i(\mathbf{b}^\sharp)[l])$ can be computed explicitly, i.e.,

$$v_s := \max_{l \in [p]} \sum_{i > h_{max}} \psi''(\mathbf{x}_i^T \mathbf{b}^\sharp)(X[i,l])^2.$$

Putting all above together, we get

$$\begin{aligned}
\|\hat{\mathbf{b}} - \mathbf{b}^\sharp\| &\le \sqrt{p}(\sqrt{v_s} + b_s)/\sigma_{min}(\nabla^2 L(\mathbf{b}^\sharp)) = O_p(\frac{\sqrt{p}(\sqrt{\psi''^{\sharp}_{max}}\|X\|_{2,\infty} + (\psi''^{\sharp}_{max}\log n + \psi'^{\sharp}_{max})h_{max})}{\sigma^2_{min}(X)}) \\
&= O_p(\frac{\sqrt{p}(\sqrt{\psi''^{\sharp}_{max}}\sqrt{n - h_{max}} + (\psi''^{\sharp}_{max}\log n + \psi'^{\sharp}_{max})h_{max})}{n\psi''^{\sharp}_{min}}\gamma_{1p}) := \delta^*.
\end{aligned} \tag{68}$$

It is clear that the bound δ^* is tighter than the first bound, δ_n.

Remark: Especially, when $\psi''^{\sharp}_{max}, \psi''^{\sharp}_{min}$ and $\psi'^{\sharp}_{max}$ are bounded and γ_{1p} is $O(1)$, then δ^* can be simplified as $\frac{\sqrt{p}(\sqrt{n-h_{max}} + (\log n)h_{max})}{n}$.

G.3. Difference of $\Lambda(\mathbf{I}, \hat{B}) - \Lambda(\Pi, \hat{B})$

Since $\Lambda(I, \hat{B})$ is the separate function of each column of $\hat{B}$, we can only focus on one column of $\hat{B}$ (denoted as $\hat{\mathbf{b}}$) in the rest of this section.

By straightforward calculation, we get

$$\begin{aligned}
& \Lambda(\mathbf{I}, \hat{\mathbf{b}}) - \Lambda(\Pi, \hat{\mathbf{b}}) \\
= \; & (\Lambda(\mathbf{I}, \hat{\mathbf{b}}) - \Lambda(\mathbf{I}, \mathbf{b}^\sharp)) - (\Lambda(\Pi, \hat{\mathbf{b}}) - \Lambda(\Pi, \mathbf{b}^\sharp)) + (\Lambda(\mathbf{I}, \mathbf{b}^\sharp) - \Lambda(\Pi, \mathbf{b}^\sharp)) \\
\ge \; & -4h_{max}\lambda^\sharp_{max}x_{max}\delta^* + \Lambda(\mathbf{I}, \mathbf{b}^\sharp) - \Lambda(\Pi, \mathbf{b}^\sharp),
\end{aligned} \tag{69}$$

where the last inequality depends on the following fact

$$\begin{aligned}
& (\Lambda(\mathbf{I}, \hat{\mathbf{b}}) - \Lambda(\mathbf{I}, \mathbf{b}^\sharp)) - (\Lambda(\Pi, \hat{\mathbf{b}}) - \Lambda(\Pi, \mathbf{b}^\sharp)) \\
= \; & (\Lambda(\mathbf{I}, \hat{\mathbf{b}}) - \Lambda(\Pi, \hat{\mathbf{b}})) - (\Lambda(\mathbf{I}, \mathbf{b}^\sharp) - \Lambda(\Pi, \mathbf{b}^\sharp)) \\
= \; & \sum_{i \le d(\mathbf{I},\Pi)} \Big\{ -\psi(\mathbf{x}_i^T\hat{\mathbf{b}}) + \psi'(\mathbf{x}_i^T\mathbf{b}^\sharp)\mathbf{x}_i^T\hat{\mathbf{b}} + \psi(\mathbf{x}_i^T\mathbf{b}^\sharp) - \psi'(\mathbf{x}_i^T\mathbf{b}^\sharp)\mathbf{x}_i^T\mathbf{b}^\sharp \\
& + \psi(\mathbf{x}_{\Pi(i)}^T\hat{\mathbf{b}}) - \psi'(\mathbf{x}_i^T\mathbf{b}^\sharp)\mathbf{x}_{\Pi(i)}^T\hat{\mathbf{b}} - \psi(\mathbf{x}_{\Pi(i)}^T\mathbf{b}^\sharp) + \psi'(\mathbf{x}_i^T\mathbf{b}^\sharp)\mathbf{x}_{\Pi(i)}^T\mathbf{b}^\sharp \Big\} \\
\le \; & \sum_{i \le d(\mathbf{I},\Pi)} 4\psi'^{\sharp}_{max} \cdot x_{max} \cdot \|\hat{\mathbf{b}} - \mathbf{b}^\sharp\| \le 4d(\mathbf{I},\Pi)\psi'^{\sharp}_{max}x_{max}\delta^*,
\end{aligned} \tag{70}$$

where $\psi'^{\sharp}_{max} = \max_i \psi'(\mathbf{x}_i^T\mathbf{b}^\sharp)$ and $x_{max} = \max_i \|\mathbf{x}_i\|$.

By (69) and summing over $l \in [m]$, we arrive at that

$$\begin{aligned}
& \Lambda(\mathbf{I}, \hat{B}) - \Lambda(\Pi, \hat{B}) \\
\ge \; & -4md(\mathbf{I},\Pi)\psi'^{\sharp}_{max}x_{max}\delta^* + \Lambda(\mathbf{I}, B^\sharp) - \Lambda(\Pi, B^\sharp), \\
\gtrsim \; & x \cdot v_{\Pi,partial},
\end{aligned} \tag{71}$$

when the conditions (14)-(15) are met, that is,

$$\Lambda(\mathbf{I}, B^\sharp) - \Lambda(\Pi, B^\sharp) \gtrsim x \cdot v_{\Pi,partial} \tag{72}$$

and

$$\Lambda(\mathbf{I}, B^\sharp) - \Lambda(\Pi, B^\sharp) \gtrsim md(\mathbf{I}, \Pi)\psi'^{\sharp}_{max} x_{max}\delta^*. \tag{73}$$

This completes the proof of Eq. (61) and Theorem 4.4 as well.

G.4. On $\Lambda(\mathbf{I}, B^\sharp) - \Lambda(\Pi, B^\sharp)$

At the end of this section, we investigate the lower bound of $\Lambda(\mathbf{I}, B^\sharp) - \Lambda(\Pi, B^\sharp)$. By elementary calculations, we have

$$\begin{aligned} & \Lambda(\mathbf{I}, B^\sharp) - \Lambda(\Pi, B^\sharp) \\ = & \sum_{l\in[m]} \sum_{i \le d(\mathbf{I},\Pi)} \Big\{ -\psi(\mathbf{x}_i^T \mathbf{b}_l^\sharp) + \psi'(\mathbf{x}_i^T \mathbf{b}_l^\sharp)\mathbf{x}_i^T \mathbf{b}_l^\sharp \\ & -(\psi(\mathbf{x}_{\Pi(i)}^T \mathbf{b}_l^\sharp) + \psi'(\mathbf{x}_i^T \mathbf{b}_l^\sharp)\mathbf{x}_{\Pi(i)}^T \mathbf{b}_l^\sharp) \Big\} \\ \ge & \sum_{l\in[m]} \sum_{i \le d(\mathbf{I},\Pi)} \frac{1}{2}\psi''^{\sharp}_{min} \big((\mathbf{x}_i^T - \mathbf{x}_{\Pi(i)}^T)\mathbf{b}_l^\sharp\big)^2. \end{aligned} \tag{74}$$

Under sub-Gaussian design, it is not hard to show $\sum_{l\in[m]} \big((\mathbf{x}_i^T - \mathbf{x}_j^T)\mathbf{b}_l^\sharp\big)^2$ is $\Omega(m)$ for any pair of $i, j \in [m]$ when $m \gtrsim \log n$ and $p \gtrsim \log n$. Thus, $\Lambda(\mathbf{I}, B^\sharp) - \Lambda(\Pi, B^\sharp)$ is $\Omega(md(\mathbf{I}, \Pi))$. Conditions (14)-(15) in Theorem 4.4 are easily satisfied.

H. Proof of Results without any knowledge of $B^\sharp$ and $\Pi^\sharp$: Theorem 4.6

In this section, we provide the proof when we do not have any knowledge of $B^\sharp$ and $\Pi^\sharp$. For each fixed permutation Π, we recall the definition,

$$v_{\Pi,B} \quad = \quad \sum_{i=1}^n \sum_{l=1}^m \psi''(\mathbf{x}_{\Pi^\sharp(i)}^T \mathbf{b}_l^\sharp)(\mathbf{x}_{\Pi(i)}^T \mathbf{b}_l)^2.$$

It can be easily checked that $v_{\Pi,B}$ is the variance of $L(\Pi, B)$.

We first compute the high probability bound of deviation, $|\langle L(\Pi, B) - \Lambda(\Pi, B)\rangle|$. Following the proof of Lemma F.1, the moment generating function of $\langle L(\Pi, B) - \Lambda(\Pi, B)\rangle$ can be upper bounded via

$$\begin{aligned} & \mathbb{E}\exp\{t\langle L(\Pi, B) - \Lambda(\Pi, B)\rangle\} \\ = & \mathbb{E}\exp\{t\sum_i \langle (Y[i,:] - \psi'(\boldsymbol{\lambda}_i^\sharp)) \circ \boldsymbol{\lambda}_{\pi_i}\rangle\} \\ = & \prod_{i=1}^n \prod_{l=1}^m \exp\{\psi(\boldsymbol{\lambda}_i^\sharp[l] + t\boldsymbol{\lambda}_i[l]) - \psi(\boldsymbol{\lambda}_i^\sharp[l]) - t\psi'(\boldsymbol{\lambda}_i^\sharp[l])\boldsymbol{\lambda}_{\pi_i}[l]\} & (75) \\ \le & \prod_{i=1}^n \prod_{l=1}^m \exp\{\psi''(\boldsymbol{\lambda}_i^\sharp[l])(\boldsymbol{\lambda}_{\pi_i}[l])^2 t^2\} & (76) \\ = & \exp\{t^2(\sum_{i,l} \psi''(\boldsymbol{\lambda}_i^\sharp[l])(\boldsymbol{\lambda}_{\pi_i}[l])^2)\} = \exp\{v_{\Pi,B}t^2\} & (77) \end{aligned}$$

for any $t \in (0, c_\psi)$. Hence we have sub-Gaussian tail bound,

$$\begin{aligned} & P(|\langle L(\Pi, B) - \Lambda(\Pi, B)\rangle| \ge v_{\Pi,B}x) \\ \le & \exp\{-v_{\Pi,B}x^2/4\} \end{aligned} \tag{78}$$

for any $x \in (0, 2c_\psi)$. By applying Markov inequality to $\mathbb{E}[\exp\{t\langle L(\Pi, B) - \Lambda(\Pi, B)\rangle\}]$ with $t = 1/\tilde{x}_{max}$, where $\tilde{x}_{max} = Cx_{max}$ for sufficiently large constant C such that $1/Cx_{max} < c_\psi$. On the other hand, we also have the following sub-exponential tail bound,

$$\begin{aligned} & P(|\langle L(\Pi, B) - \Lambda(\Pi, B)\rangle| \geq v_{\Pi,B}x) \\ \leq \; & \exp\{v_{\Pi,B}/(\tilde{x}_{max})^2\}\exp\{-v_{\Pi,B}x/\tilde{x}_{max}\} \end{aligned} \tag{79}$$

for any x.

Two situations In order to prove the results, we consider the following two situations, 1. $d(\Pi, \Pi^\sharp) \leq h_c$ 2. $d(\Pi, \Pi^\sharp) \geq h_c$ where $h_c = c_0 \frac{n}{p \log n}$. For ease of presentation, we treat $\Pi^\sharp$ as $\mathbf{I}$. (By doing this, it will not change the technical difficulty since we can always treat $\Pi(\Pi^\sharp)^{-1}$ as new Π.)

H.1. Situation 1

We first show the difference between $B(\Pi)$ and $B^\sharp$. By the definition of $B(\Pi)$, we know

$$B(\Pi) = \arg\max_B \Lambda(\Pi, B).$$

Note that $\Lambda(\Pi, B)$ is separable for each column of B, i.e.,

$$\mathbf{b}_j(\Pi) = \arg\max_{\mathbf{b}} \sum_{i=1}^n \{\psi'(\mathbf{x}_{\Pi^\sharp(i)}^T \mathbf{b}_j^\sharp)(\mathbf{x}_{\Pi(i)}^T \mathbf{b}) - \exp\{\mathbf{x}_{\Pi(i)}^T \mathbf{b}\}\}.$$

Therefore, we wish to bound the difference between $\|\mathbf{b}_j(\Pi) - \mathbf{b}_j^\sharp\|$.

By the optimality of $\mathbf{b}_j(\Pi)$, we then have that

$$\sum_{i=1}^n \psi'(\mathbf{x}_i^T \mathbf{b}_j^\sharp)(\mathbf{x}_{\Pi(i)}^T \mathbf{b}_j(\Pi)) - \psi(x_{\Pi(i)}^T \mathbf{b}_j(\Pi)) \geq \sum_{i=1}^n \psi'(\mathbf{x}_i^T \mathbf{b}_j^\sharp)(\mathbf{x}_{\Pi(i)}^T \mathbf{b}_j^\sharp) - \psi(\mathbf{x}_{\Pi(i)}^T \mathbf{b}_j^\sharp)$$

which can be written as

$$\begin{aligned} & \sum_{i:\Pi(i)\neq i} \Big\{\psi'(\mathbf{x}_i^T \mathbf{b}_j^\sharp)(\mathbf{x}_{\Pi(i)}^T \mathbf{b}_j(\Pi)) - \psi(x_{\Pi(i)}^T \mathbf{b}_j(\Pi)) - (\psi'(\mathbf{x}_i^T \mathbf{b}_j^\sharp)(\mathbf{x}_{\Pi(i)}^T \mathbf{b}_j^\sharp) - \psi(\mathbf{x}_{\Pi(i)}^T \mathbf{b}_j^\sharp))\Big\} \\ \geq \; & \sum_{i:\Pi(i)=i} \Big\{\psi'(\mathbf{x}_i^T \mathbf{b}_j^\sharp)(\mathbf{x}_{\Pi(i)}^T \mathbf{b}_j^\sharp) - \psi(\mathbf{x}_{\Pi(i)}^T \mathbf{b}_j^\sharp) - (\psi'(\mathbf{x}_i^T \mathbf{b}_j^\sharp)(\mathbf{x}_{\Pi(i)}^T \mathbf{b}_j(\Pi)) - \psi(x_{\Pi(i)}^T \mathbf{b}_j(\Pi)))\Big\}. \end{aligned} \tag{80}$$

The right hand side of (80) is bounded below by

$$\begin{aligned} RHS \; & \geq \; \frac{1}{2} \sum_{i:\Pi(i)=i} \psi''(\mathbf{x}_i^T \tilde{\mathbf{b}})(\mathbf{x}_i^T(\mathbf{b}_j^\sharp - \mathbf{b}_j(\Pi)))^2 \\ & \geq \; \frac{1}{4} cn/\gamma_{1p} \psi_{min}^{''\sharp} \|\mathbf{b}_j^\sharp - \mathbf{b}_j(\Pi)\|^2, \end{aligned} \tag{81}$$

where the last inequality depends on the curvature property which will be described later. The left hand side of (80) is bounded above by

$$\begin{aligned} LHS \; & \leq \; \sum_{i:\Pi(i)\neq i} \Big\{\psi'(\mathbf{x}_i^T \mathbf{b}_j^\sharp)(\mathbf{x}_i^T \mathbf{b}_j^\sharp) - \psi(\mathbf{x}_i^T \mathbf{b}_j^\sharp) - (\psi'(\mathbf{x}_i^T \mathbf{b}_j^\sharp)(\mathbf{x}_{\Pi(i)}^T \mathbf{b}_j^\sharp) - \psi(\mathbf{x}_{\Pi(i)}^T \mathbf{b}_j^\sharp))\Big\} \\ & \leq \; \sum_{i:\Pi(i)\neq i} (\psi'(\mathbf{x}_i^T \mathbf{b}_j^\sharp)|\mathbf{x}_i^T \mathbf{b}_j^\sharp - \mathbf{x}_{\Pi(i)}^T \mathbf{b}_j^\sharp| + \max\{\psi(\mathbf{x}_i^T \mathbf{b}_j^\sharp), \psi(\mathbf{x}_{\Pi(i)}^T \mathbf{b}_j^\sharp j)\}). \end{aligned} \tag{82}$$

Combining (81) and (82), we have

$$\|\mathbf{b}_j^\sharp - \mathbf{b}_j(\Pi)\|^2 \le C\frac{pd(\mathbf{I},\Pi)(\psi'^\sharp_{max} + \psi^\sharp_{max})\gamma_{1p}}{n\psi''^\sharp_{min}}, \tag{83}$$

by adjusting the constant C.

Given a fixed Π, we next calculate the bound of $\|\hat{B}(\Pi) - B(\Pi)\|$. By the definition (optimality) of $\hat{B}(\Pi)$ and convexity of negative log-likelihood function, we have that

$$\begin{aligned} L(\Pi, B^\sharp) \quad &\le \quad L(\Pi, \hat{B}(\Pi)) = L(\Pi, B^\sharp) + \langle \nabla L(\Pi, B^\sharp), \hat{B}(\Pi) - B^\sharp\rangle \\ &\quad + \frac{1}{2}(\hat{B}(\Pi) - B^\sharp)^T \nabla^2 L(\Pi, \tilde{B})(\hat{B}(\Pi) - B^\sharp), \end{aligned} \tag{84}$$

which gives us that

$$\begin{aligned} \frac{1}{2}(\hat{B}(\Pi) - B^\sharp)^T \nabla^2 L_{neg}(\Pi, \tilde{B})(\hat{B}(\Pi) - B^\sharp) \quad &\le \quad |\langle \nabla L(\Pi, B^\sharp), \hat{B}(\Pi) - B^\sharp\rangle| \\ \frac{1}{2}(n\psi''^\sharp_{min}/\gamma_{1p})\|\hat{B}(\Pi) - B^\sharp\|^2 \quad &\le \quad \|\hat{B}(\Pi) - B^\sharp\|\|\nabla L(\Pi, B^\sharp)\| && (85) \\ (n\psi''^\sharp_{min}/\gamma_{1p})\|\hat{B}(\Pi) - B^\sharp\|^2 \quad &\le \quad C(\psi^\sharp_{max} + \psi'^\sharp_{max})\|\hat{B}(\Pi) - B^\sharp\|\sqrt{p}(Ch\log n + \sqrt{n\log p}) && (86) \\ \|\hat{B}(\Pi) - B^\sharp\| \quad &\le \quad C(\psi^\sharp_{max} + \psi'^\sharp_{max})\frac{\gamma_{1p}\sqrt{p}(Ch\log n + \sqrt{n\log p})}{n\psi''^\sharp_{min}}, && (87) \end{aligned}$$

where we define $L_{neg}(\Pi, B) = -L(\Pi, B)$. This tells us that

$$\|\hat{B}(\Pi) - B(\Pi)\| \le \|B(\Pi) - B^\sharp\| + \|\hat{B}(\Pi) - B^\sharp\| = o_p(1)$$

as long as $ph \ll n/(\gamma_{1p}\log n)$ and $p = n^a$ ($0 < a < \frac{1}{2}$).

Curvature Property: Here (85) depends on the following observations on the curvature of log-likelihood function. Recall that Hessian matrix

$$\nabla^2 L_{neg}(\Pi, B) = (\Pi X)^T \mathrm{diag}(\psi''(\Pi X B))\Pi X \tag{88}$$

holds for any fixed permutation Π. Then

$$\begin{aligned} &(\hat{B}(\Pi) - B^\sharp)^T \nabla^2 L_{neg}(\Pi, \tilde{B})(\hat{B}(\Pi) - B^\sharp) \\ = \quad &(\hat{B}(\Pi) - B^\sharp)^T (\Pi X)^T \mathrm{diag}(\psi''(\Pi X\tilde{B}))\Pi X(\hat{B}(\Pi) - B^\sharp)^T. \end{aligned} \tag{89}$$

Let $r = \|\hat{B}(\Pi) - B^\sharp\|$ and $\mathbf{b} = \hat{B}(\Pi) - B^\sharp$. For any monotonically increasing ψ'', by assumption *A2*, we have that the cardinality of set $\mathcal{I} = \{i | \mathbf{x}_i^T\mathbf{b} \ge c_1 r\}$ is greater than n/γ_{1p}. For index i in $\mathcal{I}$, we can find that

$$\begin{aligned} \psi''(\mathbf{x}_i^T\tilde{B}) \quad &\ge \quad \min\{\psi''(\mathbf{x}_i^T\hat{B}(\Pi)), \psi''(\mathbf{x}_i^T B^\sharp)\} \\ &= \quad \min\{\psi''(\mathbf{x}_i^T(\mathbf{b} + B^\sharp)), \psi''(\mathbf{x}_i^T B^\sharp)\} \\ &\ge \quad \psi''^\sharp_{min}, \end{aligned} \tag{90}$$

since $\tilde{B}(\Pi)$ takes form of $t\hat{B}(\Pi) + (1-t)B^\sharp$. (Similarly, (90) also holds for monotonically decreasing or bounded ψ'''s.)

Thus, the right hand side of (89) can be lower bounded by

$$\begin{aligned} &(\hat{B}(\Pi) - B^\sharp)^T (\Pi X)^T \mathrm{diag}(\psi(\Pi X\tilde{B}))\Pi X(\hat{B}(\Pi) - B^\sharp) \\ \ge \quad &cn\psi''^\sharp_{min}/\gamma_{1p}r^2 \\ = \quad &cn\psi''^\sharp_{min}/\gamma_{1p}\|\hat{B}(\Pi) - B^\sharp\|^2. \end{aligned} \tag{91}$$

Similarly, we have

$$\begin{aligned} B^T\nabla^2 L_{neg}(\Pi,\bar{B})B &= B^T(\Pi X)^T \text{diag}(\psi''(\Pi X\bar{B}))\Pi X B \\ &\geq cn/\gamma_{1p}\|B\|^2 \end{aligned} \quad (92)$$

for any B and $\bar{B} = t\mathbf{0} + (1-t)B$, $(0 \leq t \leq 1)$.

We call (91) and (92) as *curvature inequalities* for log-likelihood at $B = B^\sharp$ and $B = \mathbf{0}$ correspondingly. The **most distinguished feature** of curvature inequality is that the minimum eigenvalue of Hessian matrix $\nabla^2 L_{neg}/n$ has non-trivial lower bound. That is, the eigenvalue is strictly greater than zero.

Inequality (86) comes from the following fact. For each $l \in [p]$, we consider to compute the following bound, i.e.,

$$\begin{aligned} \sup_{\Pi: d(\Pi,\mathbf{I})\leq h} |\nabla L(\Pi, B^\sharp)[l]| &\leq \sup_{\Pi: d(\Pi,\mathbf{I})\leq h} \{|\nabla L(\Pi, B^\sharp)[l] - \nabla L(\mathbf{I}, B^\sharp)[l]|\} + |\nabla L(\mathbf{I}, B^\sharp)[l]| \\ &\leq C(\psi^\sharp_{max} + \psi'^\sharp_{max})(Ch\log n + \sqrt{n\log p}), \end{aligned} \quad (93)$$

by noticing that the entry of design matrix is bounded.

In situation 1, we are going to show that

$$L(\mathbf{I}, \hat{B}) > L(\Pi, \hat{B}(\Pi))$$

with probability going to 1 as $n \to \infty$. By noticing that $L(\mathbf{I}, \hat{B}) \geq L(\mathbf{I}, \hat{B}(\Pi))$, it suffices to show

$$L(\mathbf{I}, \hat{B}(\Pi)) > L(\Pi, \hat{B}(\Pi))$$

for any Π with $d(\mathbf{I}, \Pi) \leq h_c$ with probability going to 1.

Uniform Bound of $\|\hat{B}(\Pi) - \hat{B}\|_{col}$

By (87), we then have

$$\|\hat{B}(\Pi) - \hat{B}\|_{col} \leq 2C(\psi^\sharp_{max} + \psi'^\sharp_{max})\frac{\sqrt{p}\gamma_{1p}(h_{max}\log n + \sqrt{n\log p})}{n\psi''^\sharp_{min}} =: \delta_1$$

held for any Π with $d(\Pi, \Pi^\sharp) \leq h_{max}$ by adjusting the constant C.

Difference of $|(L(\mathbf{I}, B) - L(\Pi, B)) - (\Lambda(\mathbf{I}, B) - \Lambda(\Pi, B))|$

In the following, we treat $\Pi^\sharp$ as $\mathbf{I}$ for the ease of presentation. By Lemma G.3,

$$\begin{aligned} & P(|(L(\mathbf{I}, B) - L(\Pi, B)) - (\Lambda(\mathbf{I}, B) - \Lambda(\Pi, B))| \geq 2v_{\Pi,partial}x) \\ \leq\ & \exp\{-\frac{1}{4}v_{\Pi,partial}x^2\} \end{aligned} \quad (94)$$

for any $B \in \mathcal{B}(B^\sharp, \delta)$ with $\delta = \delta_1$ and $x < c_\psi$. By using this, we can further obtain the uniform inequality, i.e.,

$$\begin{aligned} & P(\sup_{\Pi: d(\mathbf{I},\Pi)\leq h_c} \frac{1}{v_{\Pi,partial}}|(L(\mathbf{I}, \hat{B}(\Pi)) - L(\Pi, \hat{B}(\Pi))) - (\Lambda(\mathbf{I}, \hat{B}(\Pi)) - \Lambda(\Pi, \hat{B}(\Pi)))| \geq 2x) \\ \leq\ & \sum_{\Pi: d(\mathbf{I},\Pi)\leq h_c} \exp\{-\frac{1}{4}v_{\Pi,partial}x^2\} \ \text{(union bound)} \\ =\ & \sum_{h=2}^{h_c}\sum_{\Pi: d(\mathbf{I},\Pi)=h} \exp\{-\frac{1}{4}v_{\Pi,partial}x^2\} \\ \leq\ & \sum_{h=2}^{h_c} n^h\exp\{-\frac{1}{4}hv_{min}x^2\} \ \text{(using fact that } v_{\Pi,partial} \geq h\cdot v_{min}) \\ \leq\ & \frac{-2(v_{min}x^2 - \log n)}{1 - \exp\{-(v_{min}x^2 - \log n)\}}. \ \text{(summation of geometric series)} \end{aligned} \quad (95)$$

On the other hand, we could compute the difference

$$
\begin{aligned}
& |(\Lambda(\mathbf{I}, \hat{B}(\Pi)) - \Lambda(\Pi, \hat{B}(\Pi))) - (\Lambda(\mathbf{I}, B(\Pi)) - \Lambda(\Pi, B(\Pi)))|. \\
\leq & \Bigg| \sum_{i:\Pi(i)\neq i} \sum_{j=1}^{m} (\psi'(\mathbf{x}_i^T \mathbf{b}_j^\sharp)(\mathbf{x}_i^T \hat{\mathbf{b}}_j(\Pi)) - \psi(\mathbf{x}_i^T \hat{\mathbf{b}}_j(\Pi))) - (\psi'(\mathbf{x}_{\Pi(i)}^T \mathbf{b}_j^\sharp)(\mathbf{x}_i^T \hat{\mathbf{b}}_j(\Pi)) - \psi(\mathbf{x}_{\Pi(i)}^T \hat{\mathbf{b}}_j(\Pi))) \\
& - \Big((\psi'(\mathbf{x}_i^T \mathbf{b}_j^\sharp)(\mathbf{x}_i^T \mathbf{b}_j(\Pi)) - \psi(\mathbf{x}_i^T \mathbf{b}_j(\Pi))) - (\psi'(\mathbf{x}_{\Pi(i)}^T \mathbf{b}_j^\sharp)(\mathbf{x}_i^T \mathbf{b}_j(\Pi)) - \psi(\mathbf{x}_{\Pi(i)}^T \mathbf{b}_j(\Pi))) \Big) \Bigg| \\
\leq & \; 2 \sum_{i:\Pi(i)\neq i} \sum_{j=1}^{m} (\psi_{max}^\sharp + \psi_{max}^{'\sharp}) x_{max} \|\hat{\mathbf{b}}_j(\Pi) - \mathbf{b}_j(\Pi)\|. \qquad (96)
\end{aligned}
$$

Lastly, we can compute the lower bound of

$$
\begin{aligned}
& \Lambda(\mathbf{I}, B(\Pi)) - \Lambda(\Pi, B(\Pi)) \\
= & \; \Lambda(\mathbf{I}, B^\sharp) - \Lambda(\Pi, B^\sharp) - \Big(\Lambda(\mathbf{I}, B^\sharp) - \Lambda(\Pi, B^\sharp) - (\Lambda(\mathbf{I}, B(\Pi)) - \Lambda(\Pi, B(\Pi))) \Big) \\
\geq & \; \Lambda(\mathbf{I}, B^\sharp) - \Lambda(\Pi, B^\sharp) - 2 \sum_{i:\Pi(i)\neq i} \sum_{j=1}^{m} (\psi_{max}^\sharp + \psi_{max}^{'\sharp}) x_{max} \|\mathbf{b}_j^\sharp - \mathbf{b}_j(\Pi)\| \\
\geq & \; \frac{1}{2}(\Lambda(\mathbf{I}, B^\sharp) - \Lambda(\Pi, B^\sharp)) \qquad (97)
\end{aligned}
$$

by using assumption that

$$
\Lambda(\Pi^\sharp, B^\sharp) - \Lambda(\Pi, B^\sharp) \gtrsim m d(\Pi, \Pi^\sharp)(\psi_{max}^\sharp + \psi_{max}^{'\sharp}) x_{max} \delta^*.
$$

Combining (95) with x taken as $\min\{\Big(\Lambda(\mathbf{I}, B(\Pi)) - \Lambda(\Pi, B(\Pi))\Big)/2v_{\Pi,partial}, c_\psi\}$, (96) and (97), we have

$$
L(\mathbf{I}, \hat{B}(\Pi)) - L(\Pi, \hat{B}(\Pi)) > 0
$$

with probability going to 1. This implies $\hat{\Pi} = \mathbf{I} = \Pi^\sharp$.

H.2. Situation 2

In situation 2, for any fixed Π with $d(\Pi, \Pi^\sharp) \geq h_c$, we are going to bound the difference between $L(\Pi, B)$ and $\Lambda(\Pi, B)$ uniformly over all permutation matrices and the restricted parameter space.

H.2.1. On restricted space $\mathcal{B}_0$

In this section, we will first determine the restricted parameter space $\mathcal{B}_0$. First, we know that $\hat{B}(\Pi)$ is the maximizer of $L(\Pi, B)$. We have that

$$
\langle Y \circ (\Pi X \hat{B}(\Pi)) - \psi(\Pi X \hat{B}(\Pi)) \rangle \geq \langle Y \circ (\Pi X \mathbf{0}) - \psi(\Pi X \mathbf{0}) \rangle. \qquad (98)
$$

By curvature property, we have

$$
\begin{aligned}
& \langle Y \circ (\Pi X \hat{B}(\Pi)) - Y \circ (\Pi X \mathbf{0}) - \psi'(\Pi X \mathbf{0}) \circ (\Pi X \mathbf{0}) \rangle \\
\geq & \; \frac{1}{2} \psi_{min}^{''0} n \gamma_{1p} \|\hat{B}(\Pi)\|_{col}^2. \qquad (99)
\end{aligned}
$$

This implies that with high probability, it holds

$$
\|\hat{B}(\Pi)\|_{col} \leq C \frac{\sigma(X/\sqrt{n})}{\gamma_{1p} \psi_{min}^{''0}} =: \delta_{b2}, \qquad (100)
$$

since each column of $\|Y\|$ is $O_p(\sqrt{n})$. (Remark: the $\sigma(X)$ is of order $\sqrt{n}$ in many common examples, therefore δ_{b2} can be usually treated as a constant.)

Then the restricted parameter space $\mathcal{B}_0$ can be taken as

$$\mathcal{B}_0 := \{B \mid \|B\|_{col} \leq \delta_{b2}\}.$$

In other words, we know that each column of the optimizer $\hat{B}$ has the norm at most δ_{b2}.

H.2.2. Upper bound of $v_{\Pi,B}$

For any column of $B \in \mathcal{B}_0$ and $i \in [n]$, we consider to compute the upper bound of $|\mathbf{x}_i^T B|$. In fact, by Cauchy-Schwartz inequality, we have

$$|\mathbf{x}_i^T B| \leq \|\mathbf{x}_i\| \|B\| \leq x_{max}\delta_{b2}. \tag{101}$$

By the formula of $v_{\Pi,B}$, we have that

$$v_{\Pi,B} = \sum_i \psi''(\lambda_i^\sharp)(\mathbf{x}_{\pi_i}^T B)^2 \leq n\psi_{max}^{''\sharp} x_{max}^2 \delta_{b2}^2 = O(V_2). \tag{102}$$

H.2.3. Lower bound of $v_{\Pi,B}$

We consider to obtain the lower bound of $v_{\Pi,B}$ over the restricted parameter space $\mathcal{B}_0 \cap B(\mathbf{0}, \delta_{b1})^c$, where δ_{b1} is determined in (106). By the formula of $v_{\Pi,B}$, we know that

$$v_{\Pi,B} = \sum_i \psi''(\lambda_i^\sharp)(x_{\pi_i}^T B)^2. \tag{103}$$

According to assumption *A2*, we can see that there exist a constants c_a and c_b such that

$$\sharp\{i||x_{\pi_i}^T B| \geq c\|B\|\} \geq n/\gamma_{1p}.$$

Thus, we can have that

$$v_{\Pi,B} \geq c^2 n/\gamma_{1p}\|B\|^2 \psi_{min}^{''\sharp} \geq c^2 n/\gamma_{1p}\delta_{b1}^2 \psi_{min}^{''\sharp}. \tag{104}$$

H.2.4. Bound of $L(\Pi, \hat{B}(\Pi)) - \Lambda(\Pi, \hat{B}(\Pi))$

We consider two situations, $\|\hat{B}(\Pi)\|_{col} < \delta_{b1}$ (see (106)) and $\|\hat{B}(\Pi)\|_{col} \geq \delta_{b1}$. For the former one, we argue that $|L(\Pi, \hat{B}(\Pi)) - \Lambda(\Pi, \hat{B}(\Pi))|/mn$ is not far away from zero via straightforward calculations. For the latter case, we prove that $|L(\Pi, \hat{B}(\Pi)) - \Lambda(\Pi, \hat{B}(\Pi))|/mn$ also vanishes with high probability via establishing uniform concentration inequality.

When $\|\hat{B}(\Pi)\|_{col} < \delta_{b1}$, we then know that

$$\begin{aligned} & |L(\Pi, \hat{B}(\Pi)) - \Lambda(\Pi, \hat{B}(\Pi))| \\ = & |\sum_{i,l}(Y[i,l] - \psi'(\boldsymbol{\lambda}_i^\sharp[l]))(\mathbf{x}_{\Pi(i)}^T \hat{\mathbf{b}}_l)| \\ \lesssim & \log(mn)mn(\psi_{max}^{'\sharp} + \psi_{max}^{''\sharp})\sqrt{p}\delta_{b1} \end{aligned} \tag{105}$$

for any Π. We then know that

$$|L(\Pi, \hat{B}(\Pi)) - \Lambda(\Pi, \hat{B}(\Pi))| \leq K(n\log n + mp)$$

holds for some large constant K, when

$$\delta_{b1} := (n\log n + mp)/(\log(mn)mn(\psi_{max}^{'\sharp} + \psi_{max}^{''\sharp})\sqrt{p}). \tag{106}$$

When $\|\hat{B}(\Pi)\|_{col} \geq \delta_{b1}$, for any $B, B' \in \mathcal{B}_0 \cap B(\mathbf{0}, \delta_{b1})^c$ with $\|B - B'\|_{col} \leq \delta$, we have that

$$\begin{aligned} & |L(\Pi, B') - \Lambda(\Pi, B') - (L(\Pi, B) - \Lambda(\Pi, B))| \\ = & |\sum_{i,l}(Y[i,l] - \psi'(\boldsymbol{\lambda}_i^\sharp[l]))\big(\mathbf{x}_{\Pi(i)}^T \mathbf{b}_l' - \mathbf{x}_{\Pi(i)}^T \mathbf{b}_l\big)| \\ \leq & Cmn(\log n)\sqrt{p}\psi_{cb}^\sharp \delta \ \ \text{(using fact that } x\text{'s entry is bounded and } Y\text{'s entry is order of } \log n) \\ \leq & \frac{1}{2} v_{\Pi,B} x, \end{aligned} \tag{107}$$

where x in the last inequality is a fixed constant and δ can be chosen to be sufficiently small such that the lower bound of $v_{\Pi,B}x$ dominates the term $mn(\log n)\sqrt{p}\psi^{\sharp}_{cb}\delta$. It happens whenever

$$\delta \lesssim \delta_0 := x\delta_{b1}^2 \psi^{''\sharp}_{min}/(\psi^{\sharp}_{cb}(\log n)\sqrt{p}). \tag{108}$$

With (78), we then have the following sub-Gaussian uniform concentration inequality,

$$\begin{aligned} & P(\sup_{B\in\mathcal{B}_0\cap B(\mathbf{0},\delta_{b1})^c} \frac{1}{v_{\Pi,B}}|\langle L(\Pi,B)-\Lambda(\Pi,B)\rangle| \geq x_{\Pi,B}) \\ = \;& P(\sup_{B\in\mathcal{B}_g} \frac{1}{v_{\Pi,B}}|\langle L(\Pi,B)-\Lambda(\Pi,B)\rangle| \geq x_{\Pi,B}/2) \\ \leq \;& |\mathcal{B}_g| \max_{B\in\mathcal{B}_g} P(\frac{1}{v_{\Pi,B}}|\langle L(\Pi,B)-\Lambda(\Pi,B)\rangle| \geq x_{\Pi,B}/2) \\ \leq \;& |\mathcal{B}_g| \max_{B\in\mathcal{B}_g} \exp\{-v_{\Pi,B}x_{\Pi,B}^2/16\}, \end{aligned} \tag{109}$$

where $\mathcal{B}_g$ is the δ-covering net of $\mathcal{B}_0 \cap B(\mathbf{0},\delta_{b1})^c$ with $\delta \leq \delta_0$. Here we consider infinity norm on parameter space for constructing δ-covering net. Similarly, with (79), we have the following sub-exponential uniform concentration inequality,

$$\begin{aligned} & P(\sup_{B\in\mathcal{B}_0\cap B(\mathbf{0},\delta_{b1})^c} \frac{1}{v_{\Pi,B}}|\langle L(\Pi,B)-\Lambda(\Pi,B)\rangle| \geq x_{\Pi,B}) \\ \leq \;& |\mathcal{B}_g| \max_{B\in\mathcal{B}_g} \exp\{v_{\Pi,B}/(\tilde{x}_{max})^2\}\exp\{-v_{\Pi,B}x_{\Pi,B}/\tilde{x}_{max}\}. \end{aligned} \tag{110}$$

Then by straightforward calculation, the cardinality of $\mathcal{B}_g$ is bounded by $(C\frac{p}{\delta})^{mp}$ with C being some large constant.

Let $x_1 := \sqrt{(n\log n + mp\log n)v_{\Pi,B}}$, $x_2 := \max\{n\log n x_{max}, mpx_{max}\}$, and $\Delta^*_{\Pi,B} := C_1\max\{x_1,x_2\}$ (C_1 is some large constant). From (109) and (110) by taking $x_{\Pi,B} = \Delta^*_{\Pi,B}/v_{\Pi,B}$, we can obtain the uniform concentration inequality,

$$\begin{aligned} & P(\max_{\Pi\in\mathcal{P}_{large}}\sup_{B\in\mathcal{B}_0} \frac{1}{v_{\Pi,B}}|\langle L(\Pi,B)-\Lambda(\Pi,B)\rangle| \geq \Delta^*_{\Pi,B}/v_{\Pi,B}) \\ \leq \;& n!(C\frac{p}{\delta})^{mp}\max_B \min\{\exp\{-\Delta^{2}_{\Pi,B}/16v_{\Pi,B}\}, \exp\{v_{\Pi,B}/(\tilde{x}_{max})^2\}\exp\{-\Delta^*_{\Pi,B}/\tilde{x}_{max}\}\}, \\ \leq \;& \exp\{-\tilde{C}(n\log n + mp\log n)\} \\ \to \;& 0 \end{aligned} \tag{111}$$

with choice of $\delta = \frac{1}{n^2}$ in covering net and adjusting constant $\tilde{C}$.

In other words, (111) gives that

$$|L(\Pi,\hat{B}(\pi)) - \Lambda(\Pi,\hat{B}(\Pi))| = O_p(\Delta^*_{\Pi,B}). \tag{112}$$

To summarize, whenever $\|\hat{B}(\Pi)\|_{col}$ is greater than δ_{b1} or not, we always have

$$L(I,\hat{B}) - L(\Pi,\hat{B}(\Pi)) \geq \Lambda(\mathbf{I}) - \Lambda(\Pi) - O_p(\Delta^*_{\Pi,B}). \tag{113}$$

Lastly, by condition that

$$\Delta(X,B^\sharp,\Pi^\sharp,\Pi) \gtrsim \max\{\sqrt{(n+mp)mn\psi^{''\sharp}_{max}x^2_{max}\log n}, (n\log n + mp)x_{max}\}$$

for any Π satisfying $d(\Pi,\Pi^\sharp) > c_0\frac{n}{p\gamma_{1p}\log n}$ and $v_{\Pi,B} = O(mn\psi^{''\sharp}_{max}x^2_{max})$, we then have $\Delta(X,B^\sharp,\Pi,\Pi^\sharp) \gtrsim \max\{x_1,x_2\}$. Hence we get (113) is greater than 0 for all $\Pi \neq I$ satisfying $d(\mathbf{I},\Pi) > h_c$ with probability going to one. We then have $\hat{\Pi} \neq \Pi$ for any Π with $d(\mathbf{I},\Pi) > h_c$. This concludes the proof of Theorem 4.6.

H.2.5. ON ASSUMPTION *A2*

At the end of this appendix, we show that assumption *A2* is **automatically** satisfied for sub-Gaussian design setting. For simplicity, we take the Gaussian design for example, i.e., each entry of X is sampled from standard normal distribution independently. Fix $p_0 > 1/2$ and take any b with $\|b\| = 1$, find c_0 such that $\Phi(c_0) = p_0$, where $\Phi(\cdot)$ is the cumulative distribution function of standard normal random variable. Therefore,

$$|P(|\sharp\{x_i^T b > c_0\} - p_0 n| \geq nt) \leq 2\exp\{-\frac{2nt^2}{p_0(1-p_0)}\}.$$

Find ϵ-cover of unit sphere, we then have that

$$\sharp\{|\mathbf{x}_i \delta b| \leq C_1 \epsilon\} \geq n - \frac{n}{C_1},$$

for any $\|\delta b\| \leq \epsilon$. The size of ϵ-cover is bounded by $(2/\epsilon + 1)^p$. We then have that

$$P(\sharp\{\mathbf{x}_i^T \boldsymbol{\beta} > c_0 - C_1\epsilon\} \geq p_0 n - nt - n/C_1;\ \ \forall \boldsymbol{\beta}) \leq (2/\epsilon + 1)^p 2\exp\{-\frac{2nt^2}{p_0(1-p_0)}\}. \tag{114}$$

We than can choose t, p_0, C_1 and ϵ such that

$$\begin{aligned} p\log(2/\epsilon+1) &< 2nt^2/(p_0(1-p_0)) & (115)\\ p_0 n - nt - n/C_1 &\geq p & (116)\\ C_1\epsilon &= o(1). & (117) \end{aligned}$$

Then we have that

$$\sharp\{i||x_{\pi_i}^T b| \geq c_0/2\} \geq (p_0 - t - 1/C_1)n$$

with probability going to 1 as $n \to \infty$. Thus, we can see that assumption *A2* is satisfied by letting $c_1 = c_0/2$ and $\gamma_{1p} = p_0 - t - 1/C_1 = \Theta(1)$.

I. Proof of Results in the Missing Observation Case

For the purpose of completeness, we provide proof under missing observation cases. The proof strategy is similar to that of the previous setting, but computation is a bit more involved.

I.1. Size of $\mathcal{S}_l$

By the definition, we know that $\mathcal{S}_l = \{i : E[i,l] = 1\}$. Next, we give the upper and lower bound of $\mathcal{S}_l$. By Bernstein inequality, we have that

$$P(|\sum E[i,l] - qn| \geq nx) \leq \exp\{-\frac{n^2x^2}{2(nq(1-q)+nx/3)}\}. \tag{118}$$

Thus

$$P(|\sum E[i,l] - qn| \geq nq/2) \leq \exp\{-\frac{3nq}{7}\}. \tag{119}$$

In other words,

$$P(qn/2 \leq \min_l |\mathcal{S}_l| \leq \max_l |\mathcal{S}_l| \leq 3qn/2) \leq m \exp\{-\frac{3nq}{7}\}, \tag{120}$$

which means the sizes of S_l's are around qn with high probability as $nq/\log m \to \infty$. Hence, in the rest of proof, we treat that $|\mathcal{S}_l| = \Theta(qn)$.

I.2. When $B^\sharp$ is known

We first establish the following concentration lemma.

Lemma I.1. *There exists a constant c_ψ such that*

$$\begin{aligned} & P(|\langle E[i,:] \circ (\mathbf{y}_i \circ \boldsymbol{\lambda}_i - \psi(\boldsymbol{\lambda}_i))\rangle - \langle E[i,:] \circ (\mathbf{y}_i \circ \boldsymbol{\lambda}_j - \psi(\boldsymbol{\lambda}_j))\rangle - \Delta_{ij}(q)| \geq v_{ij}(q)x) \\ \leq \; & \max\{\exp\{-\frac{1}{8}v_{ij}(q)x^2\}, \exp\{-\frac{1}{8}v_{ij}(q)c_\psi^2\}\}. \end{aligned} \tag{121}$$

Proof of Lemma I.1. We calculate the variance of $\langle E[i,:] \circ (\mathbf{y}_i \circ \boldsymbol{\lambda}_i - \psi(\boldsymbol{\lambda}_i))\rangle - \langle E[i,:] \circ (\mathbf{y}_i \circ \boldsymbol{\lambda}_j - \psi(\boldsymbol{\lambda}_j))\rangle$.

$$\begin{aligned} & \mathrm{var}(\langle E[i,:] \circ (\mathbf{y}_i \circ \boldsymbol{\lambda}_i - \psi(\boldsymbol{\lambda}_i))\rangle - \langle E[i,:] \circ (\mathbf{y}_i \circ \boldsymbol{\lambda}_j - \psi(\boldsymbol{\lambda}_j))\rangle) \\ = \; & q \sum_{l=1}^m \mathrm{var}(Y[i,l](\boldsymbol{\lambda}_i[l] - \boldsymbol{\lambda}_j[l]) - (\psi(\boldsymbol{\lambda}_i[l]) - \psi(\boldsymbol{\lambda}_j[l]))) \\ & + q(1-q)(\sum_{l=1}^m \mathbb{E}[Y[i,l]](\boldsymbol{\lambda}_i[l] - \boldsymbol{\lambda}_j[l]) - (\psi(\boldsymbol{\lambda}_i[l]) - \psi(\boldsymbol{\lambda}_j[l])))^2 \\ = \; & q \sum_{l=1}^m \psi''(\boldsymbol{\lambda}_i[l])(\boldsymbol{\lambda}_i[l] - \boldsymbol{\lambda}_j[l])^2 + q(1-q) \sum_{l=1}^m (\psi'(\boldsymbol{\lambda}_i[l])(\boldsymbol{\lambda}_i[l] - \boldsymbol{\lambda}_j[l]) - (\psi(\boldsymbol{\lambda}_i[l]) - \psi(\boldsymbol{\lambda}_j[l])))^2 \\ = \; & \sum_l \{q x_{ij,2}[l] + q(1-q)(x_{ij,1}[l])^2\} \qquad (122) \\ := \; & v_{ij}(q), \qquad (123) \end{aligned}$$

where, for simplicity, we let $x_{ij,1}[l] = \psi'(\boldsymbol{\lambda}_i[l])(\boldsymbol{\lambda}_i[l] - \boldsymbol{\lambda}_j[l]) - (\psi(\boldsymbol{\lambda}_i[l]) - \psi(\boldsymbol{\lambda}_j[l]))$ and $x_{ij,2}[l] = \psi''(\boldsymbol{\lambda}_i[l])(\boldsymbol{\lambda}_i[l] - \boldsymbol{\lambda}_j[l])^2$. We may also suppress subscript i,j in $x_{ij,1}$ or $x_{ij,2}$ in the following calculations.

The moment generating function of $\langle E[i,:] \circ (\mathbf{y}_i \circ \boldsymbol{\lambda}_i - \psi(\boldsymbol{\lambda}_i))\rangle - \langle E[i,:] \circ (\mathbf{y}_i \circ \boldsymbol{\lambda}_j - \psi(\boldsymbol{\lambda}_j))\rangle - \Delta_{ij}(q)$ is

$$
\begin{aligned}
& \mathbb{E}\exp\{t(\langle E[i,:] \circ (\mathbf{y}_i \circ \boldsymbol{\lambda}_i - \psi(\boldsymbol{\lambda}_i))\rangle - \langle E[i,:] \circ (\mathbf{y}_i \circ \boldsymbol{\lambda}_j - \psi(\boldsymbol{\lambda}_j))\rangle - \Delta_{ij}(q))\} \\
= & \prod_{l=1}^{m} \mathbb{E}\exp\left\{t\Big(E[i,l](Y[i,l](\boldsymbol{\lambda}_j[l] - \boldsymbol{\lambda}_j[l]) - (\psi(\boldsymbol{\lambda}_i[l]) - \psi(\boldsymbol{\lambda}_j[l]))) - \Delta_{ij}(q)[l]\Big)\right\} \\
= & \prod_{l=1}^{m} \Big\{\big((1-q) + q\mathbb{E}\exp\{t(Y[i,l](\boldsymbol{\lambda}_i[l] - \boldsymbol{\lambda}_j[l]) - (\psi(\boldsymbol{\lambda}_i[l]) - \psi(\boldsymbol{\lambda}_j[l])))\}\big)\exp\{-t\Delta_{ij}(q)[l]\}\Big\} \\
= & \prod_{l=1}^{m} \Big\{\big((1-q) + q\exp\{\psi(\boldsymbol{\lambda}_i[l] + t(\boldsymbol{\lambda}_i[l] - \boldsymbol{\lambda}_j[l])) - \psi(\boldsymbol{\lambda}_i[l]) - t(\psi(\boldsymbol{\lambda}_i[l]) - \psi(\boldsymbol{\lambda}_j[l]))\}\big)\exp\{-t\Delta_{ij}(q)[l]\}\Big\} \\
\leq & \prod_{l=1}^{m} (1 + qx_1[l]t + aqx_2[l]t^2 + bq(x_1[l]t + ax_2[l]t^2)^2)(1 - qx_1[l]t + c(qx_1[l]t)^2) \ \ \text{(using fact-exp)} \\
\leq & \prod_{l=1}^{m} (1 + 2(aqx_2[l]t^2 + bqx_1^2[l]t^2 + cq^2x_1^2[l]t^2 - q^2x_1^2[l]t^2)) \ \ \text{(basic calculation)} \\
\leq & \prod_{l=1}^{m} (1 + 2(aqx_2[l]t^2 + bq(1-q)x_1^2[l]t^2)) \ \ \text{(basic calculation)} \\
\leq & \prod_{l=1}^{m} \exp\{2(qx_2[l] + q(1-q)x_1^2[l])t^2\} \ \ \text{(using } 1 + x \leq \exp\{x\}) \\
= & \exp\{2v_{ij}(q)t^2\},
\end{aligned} \tag{124}
$$

where we suppress symbols $x_{ij,1}, x_{ij,2}$ to x_1 and x_2 respectively. We also use the Taylor expansion for multiple times in the above inequalities which depend on the following fact,

$$\exp\{x\} \leq 1 + x + (\frac{1}{2} + \frac{1}{5}x)x^2 \ \ \text{(fact-exp)}$$

for any $|x| < 0.5$. In (124), we specifically take $a = 1, c = 1 - b, b = \frac{1}{2} + \frac{1}{10}(qx_1t + qax_2t^2) < 1$. This choice is possible and (124) holds for any $t \lesssim c'_\psi := \min_{i,j,l}\Big\{\min\{1/\boldsymbol{\lambda}_j[l], 1/(qx_1[l]), 1/\sqrt{qx_2[l]}\}\Big\}$.

Thus we have

$$
\begin{aligned}
& P(|\langle E[i,:] \circ (\mathbf{y}_i \circ \boldsymbol{\lambda}_i - \psi(\boldsymbol{\lambda}_i))\rangle - \langle E[i,:] \circ (\mathbf{y}_i \circ \boldsymbol{\lambda}_j - \psi(\boldsymbol{\lambda}_j))\rangle - \Delta_{ij}(q)| \geq v_{ij}(q)x) \\
\leq & \inf_{t \in (0, c'_\psi)} \exp\{2v_{ij}(q)t^2\}\exp\{-v_{ij}(q)xt\} \\
\leq & \max\{\exp\{-\frac{1}{8}v_{ij}(q)x^2\}, \exp\{-\frac{1}{8}v_{ij}(q)c_\psi^2\}\},
\end{aligned}
$$

with $c_\psi = 2c_{\psi'}$. □

Similar to the non-missing case when $B^\sharp$ is known, the rest of proof follows by taking the union bound over all possible pairs i and j to get the desired result.

Remark I.2. Similar to no missing observation case, we can obtain that the requirement for permutation recovery is

$$\Delta_{ij}(q) \gtrsim \sqrt{(\log n)v_{ij}(q)} \ \ \text{and} \ \ v_{ij}(q) \gtrsim \log n.$$

That is,

$$\Delta_{ij}(q)^2 \gtrsim q\sum_{l=1}^{m} \psi''(\boldsymbol{\lambda}_i[l])(\boldsymbol{\lambda}_i[l] - \boldsymbol{\lambda}_j[l])^2 + q(1-q)\sum_{l=1}^{m}(\psi'(\boldsymbol{\lambda}_i[l])(\boldsymbol{\lambda}_i[l] - \boldsymbol{\lambda}_j[l]) - (\psi(\boldsymbol{\lambda}_i[l]) - \psi(\boldsymbol{\lambda}_j[l])))^2 \gtrsim \log n.$$

Especially, when $\lambda_{il}^\sharp$'s are bounded and $\min_{i,j}\sum_{l\in[m]}(\lambda_{il}^\sharp - \lambda_{jl}^\sharp)^2 = \Omega(m)$, the above inequality becomes

$$q^2m^2 \gtrsim \log n(qm + q(1-q)m) \ \ \text{and} \ \ qm \gtrsim \log n.$$

Thus $q \geq \frac{\log n}{m}$ is required for the perfect permutation recovery.

I.3. With knowledge that $d(\mathbf{I}, \Pi^\sharp)$ is small

Again, in order to prove the recovery consistency, we need to control the following quantities, $\|B - B^\sharp\|_{col}$ and $\sup_{B \in B_\delta(B^\sharp)} |L(\mathbf{I}, B, E) - L(\Pi, B, E)) - (\Lambda(\mathbf{I}, B, q) - \Lambda(\Pi, B, q)|$. For ease of presentation, we still treat $\Pi^\sharp = \mathbf{I}$ here.

Suppose we have already known that the estimator $\hat{B}$ which is close to the truth $B^\sharp$, i.e., in the δ-neighborhood of $B^\sharp$. Then we let $B_\delta(B^\sharp) := \{B : \|B - B^\sharp\|_{col} \leq \delta\}$ and δ is a sufficiently small constant which will be determined later. By simple modifications of Lemma G.3 and Lemma I.1, for any fixed Π, we have

$$\begin{aligned}
& P(\sup_{B \in B_\delta(B^\sharp)} |(L(\mathbf{I}, B, E) - L(\Pi, B, E)) - (\Lambda(\mathbf{I}, B, q) - \Lambda(\Pi, B, q))| \geq 2v_{\Pi,partial,q}x) \\
\leq\ & P(|(L(\mathbf{I}, B^\sharp, E) - L(\Pi, B^\sharp, E)) - (\Lambda(\mathbf{I}, B^\sharp, q) - \Lambda(\Pi, B^\sharp, q))| \geq v_{\Pi,partial,q}x) \\
\leq\ & \exp\{-\frac{1}{4}v_{\Pi,partial,q}x^2\}
\end{aligned} \tag{125}$$

for $x \leq c_\psi$.

By calculations, we get

$$\begin{aligned}
& P(\max_\Pi \sup_{B \in B_\delta(B^\sharp)} |(L(\mathbf{I}, B, E) - L(\Pi, B, E)) - (\Lambda(\mathbf{I}, B, q) - \Lambda(\Pi, B, q))| \geq 2v_{\Pi,partial,q}x) \\
\leq\ & \sum_\Pi P(|(L(\mathbf{I}, B^\sharp) - L(\Pi, B^\sharp)) - (\Lambda(\mathbf{I}, B^\sharp) - \Lambda(\Pi, B^\sharp))| \geq v_{\Pi,partial,q}x) \\
=\ & \sum_h \sum_{\Pi:d(\Pi,\mathbf{I})=h} \cdot \exp\{-v_{\Pi,partial,q}x^2/4\} \\
\leq\ & \sum_h n!/(n-h)! \cdot \exp\{-v_{\Pi,partial,q}x^2/4\} \ \text{(basic calculation)} \\
\leq\ & \sum_h n^h \cdot \exp\{-hv_{min,q}x^2/4\} \ \text{(basic calculation)} \\
=\ & \sum_h \exp\{-h(v_{min,q}x^2 - \log n)\} \ \text{(using } v_{\Pi,partial,q} \geq h \cdot v_{min,q}) \\
\leq\ & \frac{\exp\{-2(v_{min,q}x^2 - \log n)\}}{1 - \exp\{-(v_{min,q}x^2 - \log n)\}},
\end{aligned} \tag{126}$$

where we recall that $v_{min,q} = \min_{i,j} \sum_l \left\{ q\psi''(\boldsymbol{\lambda}_i^\sharp[l])(\boldsymbol{\lambda}_j^\sharp[l] - \boldsymbol{\lambda}_i^\sharp[l])^2 + q(1-q)(\psi'(\boldsymbol{\lambda}_i^\sharp[l])(\boldsymbol{\lambda}_i^\sharp[l] - \boldsymbol{\lambda}_j^\sharp[l]) - (\psi(\boldsymbol{\lambda}_i^\sharp[l]) - \psi(\boldsymbol{\lambda}_k^\sharp[l])))^2 \right\}$.

By (126), with probability going to 1, we have that $L(\Pi, \hat{B}, E) \leq L(\mathbf{I}, \hat{B}, E) - \Lambda(\mathbf{I}, \hat{B}, q) + \Lambda(\Pi, \hat{B}, q) + x \cdot v_{\Pi,B^\sharp,q}$ with $x = \min\{C\sqrt{\log n / v_{min}}, c_\psi\}$ with C being some constant. This tells us that if we can show that

$$\Lambda(\mathbf{I}, \hat{B}, q) - \Lambda(\Pi, \hat{B}, q) - x \cdot v_{\Pi,B^\sharp,q} > 0. \tag{127}$$

Then we can conclude that $\hat{\Pi} = \mathbf{I}$ which leads to the desired result.

I.3.1. First bound of $\|B - B^\sharp\|_{col}$

Again we first show that $\|B - B^\sharp\|_{col} = o_p(1)$. Column-wisely, we need to construct a δ_n such that for any $\mathbf{b}$ with $\|\mathbf{b} - \mathbf{b}^\sharp\| \geq \delta_n$, it holds $L(\mathbf{b}) < L(\mathbf{b}^\sharp)$. By the definition that $L(\hat{\mathbf{b}}) \geq L(\mathbf{b}^\sharp)$, we will arrive at $\|\hat{\mathbf{b}} - \mathbf{b}^\sharp\| \leq \delta_n$.

Similar to the non-missing case, we can take

$$\delta_n^2 := C\frac{\sqrt{v_{b^\sharp,q}} + q\lambda_{max}^\sharp h_{max}}{q(n - h_{max})\lambda_{min}^\sharp / \gamma_{2p}}$$

with some constant C and

$$v_{b,q} := \sum_{i=1}^{n}\{q\psi''(\boldsymbol{\lambda}_i^\sharp)(\boldsymbol{\lambda}_i(\mathbf{b}))^2 + q(1-q)\big(\psi'(\boldsymbol{\lambda}_i^\sharp)\boldsymbol{\lambda}_i(\mathbf{b}) - \psi(\boldsymbol{\lambda}_i(\mathbf{b}))\big)\}.$$

With this choice of δ_n, we can check that $\delta_n = o_p(1)$ and

$$L(\mathbf{b}^\sharp) - L(\mathbf{b}) > 0$$

holds for any $\mathbf{b}$ with $\|\mathbf{b} - \mathbf{b}^\sharp\| = \delta_n$ when $p^2/q < n$ and $ph_{max} < n/\log n$. By the concavity of likelihood function, we then know that $\|\hat{\mathbf{b}} - \mathbf{b}^\sharp\| \leq \delta_n$.

I.3.2. SECOND BOUND OF $\|B - B^\sharp\|_{col}$

We do the Taylor expansion of $L(\mathbf{b})$ at $\mathbf{b} = \mathbf{b}^\sharp$. Then, it can be computed that

$$\mathbf{0} = \nabla L(\hat{\mathbf{b}}) = \nabla L(\mathbf{b}^\sharp) + \nabla^2 L(\bar{\mathbf{b}})(\hat{\mathbf{b}} - \mathbf{b}^\sharp)$$

where $\bar{\mathbf{b}}$ is some point between $\hat{\mathbf{b}}$ and $\mathbf{b}^\sharp$. Again, by curvature inequality technique under assumption *E2*, we have

$$\begin{aligned} & \|\nabla^2 L(\bar{\mathbf{b}})(\hat{\mathbf{b}} - \mathbf{b}^\sharp)\| \\ \geq \quad & cqn\psi_{min}^{''\sharp}/\gamma_{2p}\|\hat{\mathbf{b}} - \mathbf{b}^\sharp\|. \end{aligned} \tag{128}$$

We thus have

$$\|\hat{\mathbf{b}} - \mathbf{b}^\sharp\| = \|(\nabla^2 L(\bar{\mathbf{b}}))^{-1}\nabla L(\mathbf{b}^\sharp)\| \leq \frac{1}{cqn\psi_{min}^{''\sharp}/\gamma_{2p}}\|\nabla L(\mathbf{b}^\sharp)\|. \tag{129}$$

For l-th element of $\nabla L(\mathbf{b}^\sharp)$, we can find that

$$\nabla L(\mathbf{b}^\sharp)[l] = \sum_{i \leq h_{max}: E[i]=1} \nabla L_i(\mathbf{b}^\sharp)[l] + \sum_{i > h_{max}: E[i]=1} \nabla L_i(\mathbf{b}^\sharp)[l].$$

The first term is bounded by

$$b_s := |\sum_{i \leq h_{max}} (Y[i] - \psi'(\mathbf{x}_i^T\mathbf{b}^\sharp))X[i,l]\}|,$$

which is order of $(\log n)\sum_{i \leq h_{max}} \psi_{max}^{''\sharp}|X[i,l]|$. The second term is bounded by $C\sqrt{\text{var}(\sum_{i>h_{max}} \nabla L_i(\mathbf{b}^\sharp)[l])}$ and the upper bound of $\text{var}(\sum_{i>h_{max}:E[i]=1} \nabla L_i(\mathbf{b}^\sharp)[l])$ can be computed explicitly, i.e.,

$$v_{s,q} := \max_{l\in[p]} \sum_{i>h_{max}} q\psi''(\mathbf{x}_i^T\mathbf{b}^\sharp)(X[i,l])^2 + q(1-q)(\psi'(\mathbf{x}_i^T\mathbf{b}^\sharp)X[i,l] - \psi(\mathbf{x}_i^T\mathbf{b}^\sharp))^2.$$

Thus

$$\begin{aligned} \|\hat{\mathbf{b}} - \mathbf{b}^\sharp\| \quad & \leq \quad \sqrt{p}(\sqrt{v_{s,q}} + b_s)/(cqn\psi_{min}^{''\sharp}/\gamma_{2p}) = O_p(\frac{\sqrt{p}(\sqrt{v_{s,q}} + \psi_{max}^{''\sharp}h_{max}\log n)}{qn}) \\ & = \quad O_p(\frac{\sqrt{p}(\sqrt{q\psi_{max}^{''\sharp} + q(1-q)\psi_{cb}^{\sharp 2}}\sqrt{n - h_{max}} + \psi_{max}^{''\sharp}h_{max}\log n)}{qn\psi_{min}^{''\sharp}/\gamma_{2p}}) := \delta_q^*. \end{aligned} \tag{130}$$

Thus the bound δ^* is tighter than δ_n. Especially, when $\mathbf{x}_i^T\mathbf{b}^\sharp$ is bounded for all i and $\gamma_{2p} = O(1)$, then δ^* can be simplified as $\frac{\sqrt{p}(\sqrt{q(n-h_{max})} + h_{max}\log n)}{qn}$.

I.3.3. Difference of $\Lambda(\mathbf{I}, \hat{\mathbf{b}}, q) - \Lambda(\Pi, \hat{\mathbf{b}}, q)$

By straightforward calculation, we get

$$\begin{aligned}
& \Lambda(\mathbf{I}, \hat{\mathbf{b}}, q) - \Lambda(\Pi, \hat{\mathbf{b}}, q) \\
= \; & q\Big\{(\Lambda(\mathbf{I}, \hat{\mathbf{b}}) - \Lambda(\mathbf{I}, \mathbf{b}^{\sharp})) - (\Lambda(\Pi, \hat{\mathbf{b}}) - \Lambda(\Pi, \mathbf{b}^{\sharp})) + (\Lambda(\mathbf{I}, \mathbf{b}^{\sharp}) - \Lambda(\Pi, \mathbf{b}^{\sharp}))\Big\} \\
\geq \; & -4qh_{max}\psi^{\sharp}_{cb}x_{max}\delta^{*} + q(\Lambda(\mathbf{I}, \mathbf{b}^{\sharp}) - \Lambda(\Pi, \mathbf{b}^{\sharp})), \qquad (131)
\end{aligned}$$

where the last inequality holds due to the same reason as explained in no-missing observation case. By (131) and summing over $l \in [m]$, we have that

$$\begin{aligned}
& \Lambda(\mathbf{I}, \hat{B}, q) - \Lambda(\Pi, \hat{B}, q) \\
\geq \; & -4qmd(\mathbf{I}, \Pi)\psi^{\sharp}_{cb}x_{max}\delta^{*} + q(\Lambda(\mathbf{I}, \mathbf{b}^{\sharp}) - \Lambda(\Pi, \mathbf{b}^{\sharp})), \\
\gtrsim \; & xv_{\Pi, partial, q}, \qquad (132)
\end{aligned}$$

when

$$q(\Lambda(\mathbf{I}, \mathbf{b}^{\sharp}) - \Lambda(\Pi, \mathbf{b}^{\sharp})) \gtrsim xv_{\Pi, partial, q}$$

and

$$\Lambda(\mathbf{I}, \mathbf{b}^{\sharp}) - \Lambda(\Pi, \mathbf{b}^{\sharp}) \gtrsim md(\mathbf{I}, \Pi)\psi^{\sharp}_{cb}x_{max}\delta^{*}.$$

This completes the proof if we take $x = \sqrt{\log n / v_{min,q}}$.

I.4. With no knowledge of $B^{\sharp}$ and $\Pi^{\sharp}$

We consider to compute the moment generating function of $\langle L(\Pi, B, E) - \mathbb{E}\Lambda(\Pi, B, E)\rangle$. Similar to Lemma I.1, we can obtain

$$\begin{aligned}
& \mathbb{E}\exp\{t\langle L(\Pi, B, E) - \mathbb{E}\Lambda(\Pi, B, E)\rangle\} \\
\leq \; & \prod_{i=1}^{n}\prod_{l=1}^{m}\mathbb{E}\exp\left\{t\Big(E[i,l](Y[i,l]\boldsymbol{\lambda}_{\Pi(i)}[l] - \psi(\boldsymbol{\lambda}_{\Pi(i)}[l])) - q(\psi'(\boldsymbol{\lambda}^{\sharp}_{i}[l])\boldsymbol{\lambda}_{\Pi(i)}[l] - \psi(\boldsymbol{\lambda}_{\Pi(i)}[l]))\Big)\right\} \\
\leq \; & \exp\{2v_{\Pi, B, q}t^2\}
\end{aligned}$$

for $t \leq \frac{c}{\psi'(x_{max}\sqrt{p/q}) + \psi''(x_{max}\sqrt{p/q})} =: 1/g(n,p)$ (c is some small constant, $g(n,p)$ is around of order $\psi^{\sharp}_{cb}$), where

$$v_{\Pi, B, q} = q\sum_{i=1}^{n}\sum_{l=1}^{m}\psi''(\boldsymbol{\lambda}^{\sharp}_{i}[l])(\boldsymbol{\lambda}_{\Pi(i)}[l])^2 + q(1-q)\sum_{i=1}^{n}\sum_{l=1}^{m}(\psi'(\boldsymbol{\lambda}^{\sharp}_{i}[l])\boldsymbol{\lambda}_{\Pi}[l] - \psi(\boldsymbol{\lambda}_{\Pi(i)}[l]))^2.$$

Thus we have sub-Gaussian tail probability,

$$\begin{aligned}
& P(|\langle L(\Pi, B) - \Lambda(\Pi, B)\rangle| \geq v_{\Pi, B, q}x) \\
\leq \; & \exp\{-v_{\Pi, B, q}x^2/4\}
\end{aligned}$$

for any $x \leq \frac{2}{g(n,p)}$. We also have the following exponential tail probability,

$$\begin{aligned}
& P(|\langle L(\Pi, B) - \Lambda(\Pi, B)\rangle| \geq v_{\Pi, B, q}x) \\
\leq \; & \exp\{v_{\Pi, B, q}/(g(n,p))^2\}\exp\{-v_{\Pi, B, q}x/g(n,p)\}
\end{aligned}$$

for any x.

By the same logic, we still consider two situations, 1. $d(\Pi, \Pi^\sharp) \le h_c$ 2. $d(\Pi, \Pi^\sharp) \ge h_c$ where $h_c = c_0 \frac{nq}{p\gamma_{3p} \log n}$. We first show the difference between $B(\Pi)$ and $B^\sharp$. By the definition of $B(\Pi)$, we know

$$B(\Pi) = \arg\max_B \Lambda(\Pi, B, q).$$

Note that $\Lambda(\Pi, B, q)$ is also separable for each column of B, i.e.,

$$\mathbf{b}_j(\Pi) = \arg\max_{\mathbf{b}} q \Big\{ \sum_{i=1}^n \psi'(\mathbf{x}_i^T \mathbf{b}_j^\sharp)(\mathbf{x}_{\Pi(i)}^T \mathbf{b}) - \psi(\mathbf{x}_{\Pi(i)}^T \mathbf{b}) \Big\}.$$

Here we still assume $\Pi^\sharp = \mathbf{I}$ without loss of generality. Notice that that the maximizer of $\Lambda(\Pi, B, q)$ remains the same as that of $\Lambda(\Pi, B)$. Thus, the difference $\|\mathbf{b}_j(\Pi) - \mathbf{b}_j^\sharp\|$ has already been obtained as before.

I.4.1. Situation 1

For situation 1, we are going to show

$$L(\mathbf{I}, \hat{B}, E) > L(\Pi, \hat{B}(\Pi), E)$$

with high probability. It suffices to show

$$L(\mathbf{I}, \hat{B}(\Pi), E) > L(\Pi, \hat{B}(\Pi), E)$$

for any Π with $d(\mathbf{I}, \Pi) \le h_c$.

Given Π, we aim to calculate the bound of $\|\hat{B}(\Pi) - B(\Pi)\|_{col}$. By the definition of $\hat{B}(\Pi)$ and convexity of negative log-likelihood function, we have that

$$\begin{aligned} L(\Pi, B^\sharp, E) \ge L(\Pi, \hat{B}(\Pi), E) &= L(\Pi, B^\sharp, E) + \langle \partial L(\Pi, B^\sharp, E), \hat{B}(\Pi) - B^\sharp \rangle \\ &\quad + \frac{1}{2} (\hat{B}(\Pi) - B^\sharp)^T \nabla^2 L(\Pi, \tilde{B}, E)(\hat{B}(\Pi) - B^\sharp), \end{aligned}$$

which gives us that

$$\begin{aligned} \frac{1}{2} (\hat{B}(\Pi) - B^\sharp)^T \nabla^2 L(\Pi, \tilde{B}, E)(\hat{B}(\Pi) - B^\sharp) &\le |\langle \nabla L(\Pi, B^\sharp, E), \hat{B}(\Pi) - B^\sharp \rangle| \\ \frac{1}{2} (nq\psi''^\sharp_{min}/\gamma_{3p}) \|\hat{B}(\Pi) - B^\sharp\|^2 &\le \|\hat{B}(\Pi) - B^\sharp\| \|\langle \nabla L(\Pi, B^\sharp, E)\| \\ (nq\psi''^\sharp_{min}/\gamma_{3p}) \|\hat{B}(\Pi) - B^\sharp\|^2 &\le C\psi^\sharp_{cb} \|\hat{B}(\Pi) - B^\sharp\| \sqrt{p}(Ch \log n + \sqrt{n \log p}) \end{aligned} \tag{133}$$

$$\|\hat{B}(\Pi) - B^\sharp\| \le C\psi^\sharp_{cb} \frac{\sqrt{p}(Ch \log n + \sqrt{nq \log p})}{\psi''^\sharp_{min} nq/\gamma_{3p}}. \tag{134}$$

This tells us that

$$\|\hat{B}(\Pi) - B(\Pi)\| \le \|B(\Pi) - B^\sharp\| + \|\hat{B}(\Pi) - B^\sharp\| = o(1)$$

as long as $ph \ll qn/(\gamma_{3p} \log n)$ and $p = (qn)^{1/2 - o(1)}$.

Here, (133) comes from the following fact. For each $l \in [p]$, we consider to compute the following bound, i.e.,

$$\begin{aligned} \sup_{\Pi: d(\Pi, \mathbf{I}) \le h} |\nabla L(\Pi, B^\sharp, E)[l]| &\le \sup_{\Pi: d(\Pi, \mathbf{I}) \le h} \{ |\nabla L(\Pi, B^\sharp, E)[l] - \nabla L(\mathbf{I}, B^\sharp, E)[l]| \} + |\nabla L(\mathbf{I}, B^\sharp, E)[l]| \\ &\le C\psi^\sharp_{cb}(Ch \log n + \sqrt{nq \log p}), \end{aligned} \tag{135}$$

by noticing that the entry of design matrix is bounded.

Uniform Bound of $\|\hat{B}(\Pi) - \hat{B}\|_{col}$ By (134), we have

$$\|\hat{B}(\Pi) - \hat{B}\|_{col} \le 2C\psi^\sharp_{cb} \frac{\sqrt{p}(h \log n + \sqrt{nq \log p})}{\psi''^\sharp_{min} nq/\gamma_{3p}} =: \delta_2^*$$

by adjusting the constant.

Uniform Concentration Inequality Similar to non-missing case, we can obtain that

$$\begin{aligned} & P(|(L(\mathbf{I}, B, E) - L(\Pi, B, E)) - (\Lambda(\mathbf{I}, B, q) - \Lambda(\Pi, B, q))| \geq v_{\Pi,partial,q} x) \\ \leq \; & \exp\{-\frac{1}{4} v_{\Pi,partial,q} x^2\} \end{aligned} \tag{136}$$

for any fixed $B \in \mathcal{B}(B^\sharp, \delta_2^*)$ and suitable x. By using this, we can further obtain the uniform inequality:

$$\begin{aligned} & P(\sup_{\Pi: d(\mathbf{I},\Pi) \leq h_c} \frac{1}{v_{\Pi,partial,q}} |(L(\mathbf{I}, \hat{B}(\Pi), E) - L(\Pi, \hat{B}(\Pi), E) - (\Lambda(\mathbf{I}, \hat{B}(\Pi), q) - \Lambda(\Pi, \hat{B}(\Pi), q))| \geq x) \\ \leq \; & \sum_{\Pi: d(\mathbf{I},\Pi) \leq h_c} \exp\{-\frac{1}{4} v_{\Pi,partial,q} x^2\} \\ \leq \; & \sum_{h=2}^{h_c} \sum_{\Pi: d(\mathbf{I},\Pi) = h} \exp\{-\frac{1}{4} v_{\Pi,partial,q} x^2\} \\ \leq \; & \sum_{h=2}^{h_c} n^h \exp\{-\frac{1}{4} h v_{min,q} x^2\} \\ \leq \; & \frac{-2(v_{min,q} x^2 - \log n)}{1 - \exp\{-(v_{min,q} x^2 - \log n)\}}. \end{aligned} \tag{137}$$

On the other hand, we calculate the difference

$$\begin{aligned} & |(\Lambda(\mathbf{I}, \hat{B}(\Pi), q) - \Lambda(\Pi, \hat{B}(\Pi), q) - (\Lambda(\mathbf{I}, B(\Pi), q) - \Lambda(\Pi, B(\Pi), q))| \\ \leq \; & q \Big| \sum_{i:\Pi(i) \neq i} \sum_{j=1}^m (\psi'(\mathbf{x}_i^T \mathbf{b}_j^\sharp)(\mathbf{x}_i^T \hat{\mathbf{b}}_j(\Pi)) - \psi(\mathbf{x}_i^T \hat{\mathbf{b}}_j(\Pi))) - (\psi'(\mathbf{x}_{\Pi(i)}^T \mathbf{b}_j^\sharp)(\mathbf{x}_i^T \hat{\mathbf{b}}_j(\Pi)) - \psi(\mathbf{x}_{\Pi(i)}^T \hat{\mathbf{b}}_j(\Pi))) \\ & - \Big((\psi'(\mathbf{x}_i^T \mathbf{b}_j^\sharp)(\mathbf{x}_i^T \mathbf{b}_j(\Pi)) - \psi(\mathbf{x}_i^T \mathbf{b}_j(\Pi))) - (\psi'(\mathbf{x}_{\Pi(i)}^T \mathbf{b}_j^\sharp)(\mathbf{x}_i^T \mathbf{b}_j(\Pi)) - \psi(\mathbf{x}_{\Pi(i)}^T \mathbf{b}_j(\Pi))) \Big) \Big| \\ \leq \; & 2q \sum_{i:\Pi(i) \neq i} \sum_{j=1}^m \psi_{cb}^\sharp x_{max} \|\hat{\mathbf{b}}_j(\Pi) - \mathbf{b}_j(\Pi)\|. \end{aligned} \tag{138}$$

Moreover, we can compute the

$$\begin{aligned} & \Lambda(\mathbf{I}, B(\Pi), q) - \Lambda(\Pi, B(\Pi), q) \\ = \; & \Lambda(\mathbf{I}, B^\sharp, q) - \Lambda(\Pi, B^\sharp, q) - \Big(\Lambda(\mathbf{I}, B^\sharp, q) - \Lambda(\Pi, B^\sharp, q) - (\Lambda(\mathbf{I}, B(\Pi), q) - \Lambda(\Pi, B(\Pi)), q) \Big) \\ \geq \; & \Lambda(\mathbf{I}, B^\sharp, q) - \Lambda(\Pi, B^\sharp, q) - 2 \sum_{i:\Pi(i) \neq i} \sum_{j=1}^m \psi_{cb}^\sharp x_{max} \|\mathbf{b}_j^\sharp - \mathbf{b}_j(\Pi)\| \\ \geq \; & c(\Lambda(\mathbf{I}, B^\sharp, q) - \Lambda(\Pi, B^\sharp, q), . \end{aligned} \tag{139}$$

Combining (137), (138) with x taken as $\Lambda(\mathbf{I}, B(\Pi), q) - \Lambda(\Pi, B(\Pi), q)/2v_{\Pi,partial,q}$ and (139), we have

$$L(\mathbf{I}, \hat{B}(\Pi), E) - L(\Pi, \hat{B}(\Pi), E) \geq c(\Lambda(\mathbf{I}, B^\sharp, q) - \Lambda(\Pi, B^\sharp, q)) > 0$$

with probability going to 1.

I.4.2. SITUATION 2

In situation 2, for any Π with $d(\Pi, \Pi^\natural) \geq h_c$ and parameter matrix B, we are going to show the high probability bound of $|\langle L(\Pi, B, E) - \Lambda(\Pi, B, q)\rangle|$. Then bound of $|\langle L(\Pi, \hat{B}(\Pi), E) - \Lambda(\Pi, \hat{B}(\Pi), q)\rangle|$ follows as well. The main proof strategy is similar to that in non-missing observation setting.

On $\mathcal{B}_0$ In this part, we first determined the restricted parameter space $\mathcal{B}_0$. First, we know that $\hat{B}(\Pi)$ is the maximizer of $L(\Pi, B, E)$. We let $\nabla L(\Pi, B, E) = (\Pi X)^T(E \circ Y) - (\Pi X)^T(E \circ \psi'(\Pi X B))$ and then know that

$$\nabla L(\Pi, \hat{B}(\Pi), E) = \mathbf{0} \tag{140}$$

and

$$\nabla L(\Pi, \mathbf{0}, E) = (\Pi X)^T(E \circ Y) - (\Pi X)^T(E \circ \psi'(\mathbf{1})). \tag{141}$$

By Talyor expansion of $\nabla L(\Pi, B, E)$, we have that

$$\nabla L(\Pi, \mathbf{0}, E) = \nabla L(\Pi, \hat{B}(\Pi), E) + \nabla^2 L(\Pi, \tilde{B}(\Pi), E)\hat{B}(\Pi). \tag{142}$$

By the formula that $\nabla^2 L(\Pi, B, E) = \Pi^T X^T \text{diag}(E \circ \psi''(\Pi X B))\Pi X$, we can easily obtain that

$$\|\hat{B}(\Pi)\|_{col} \leq C\sqrt{p}(\log n)\gamma_{3p}\psi^\natural_{cb}/(x_{max}\psi^{''0}_{min}) := \delta^{'}_{b2}.$$

Then the restricted parameter space $\mathcal{B}_0$ is taken as

$$\mathcal{B}_0 := \{B \,| \|B\|_{col} \leq \delta^{'}_{b2}\}.$$

Upper bound of $v_{\Pi,B,q}$ We need to estimate the upper bound of $v_{\Pi,B}$ over the restricted parameter space. By the formula of $v_{\Pi,B,q}$, we have that

$$\begin{aligned} v_{\Pi,B,q} &= q\sum_i \psi''(\lambda^\natural_i)(\mathbf{x}^T_{\Pi(i)}B)^2 + q(1-q)\sum_i(\psi'(\lambda^\natural_i)(\mathbf{x}^T_{\Pi(i)}B) - \psi(\mathbf{x}^T_{\Pi(i)}B))^2 \\ &\leq qn\psi^{''\natural}_{max}(x_{max}\delta'_{b2})^2 \\ &\quad +2q(1-q)(\psi^{'\natural}_{max}x_{max}\delta'_{b2})^2 + 2q(1-q)(\psi(x_{max}\delta'_{b2}))^2 := V_2(q) \end{aligned} \tag{143}$$

Lower bound of $v_{\Pi,B,q}$ We consider to obtain the lower bound of $v_{\Pi,B,q}$ over the restricted parameter space $\mathcal{B}_0$. By the formula of $v_{\Pi,B,q}$, we know that

$$v_{\Pi,B,q} = q\sum_i \psi''(\lambda^\natural_i)(\mathbf{x}^T_{\Pi(i)}B)^2 + q(1-q)\sum_i(\psi'(\lambda^\natural_i)(\mathbf{x}^T_{\Pi(i)}B) - \psi(\mathbf{x}^T_{\Pi(i)}B))^2. \tag{144}$$

Let $\delta_b := \|B\|$ and we have Then we will have

$$v_{\Pi,B,q} \geq qc/\gamma_{3p}n\psi^{''\natural}_{min}\delta^2_b.$$

For any B with $\delta_b \leq d_0/(x_{max}\max\{1, \psi^{'\natural}_{max}\})$ with d_0 satisfying $|\psi(x)| > d_0/2$ for any $|x| < d_0$, it then holds

$$\sum_i(\psi'(\lambda^\natural_i)(\mathbf{x}^T_{\Pi(i)}B) - \psi(\mathbf{x}^T_{\Pi(i)}B))^2 \geq nd_0^2/4.$$

Therefore, $v_{\Pi,B,q} \geq \min\{cqn\psi^{''\natural}_{min}\delta^2_{b1}/\gamma_{3p}, q(1-q)n/4\} := v_{lb,q}$, where $\delta_{b1} := d_0/(x_{max}\max\{1, \psi^{'\natural}_{max}\})$.

Bound of $|\langle L(\Pi, B, E) - \Lambda(\Pi, B, q)\rangle|$

Define $\Delta^*_{\Pi,B,q} := C_2 \max\{\sqrt{(n\log n + mp\log(n))v_{\Pi,B,q}}, (n\log n + mp)(\log n)g(n,p)\}$ with C_2 being a large constant. Similar to non-missing observation case, we can obtain the following uniform sub-Gaussian concentration inequality,

$$\begin{aligned} & P(\max_{\Pi} \sup_{B\in\mathcal{B}_0} \frac{1}{v_{\Pi,B,q}} |\langle L(\Pi,B,E) - \Lambda(\Pi,B,q)\rangle| \geq \Delta^*_{\Pi,B,q}/v_{\Pi,B,q}) \\ \leq \quad & n! \sum_{B\in\mathcal{B}_g} \exp\{-\Delta^*_{\Pi,B,q} x^2/16 v_{\Pi,B,q}\}, \end{aligned} \tag{145}$$

and also obtain the following uniform sub-exponential concentration inequality,

$$\begin{aligned} & P(\max_{\Pi} \sup_{B\in\mathcal{B}_0} \frac{1}{v_{\Pi,B,q}} |\langle L(\Pi,B,E) - \Lambda(\Pi,B,q)\rangle| \geq \Delta^*_{\Pi,B,q}/v_{\Pi,B,q}) \\ \leq \quad & n! \sum_{B\in\mathcal{B}_g} \exp\{v_{\Pi,B,q}/(g(n,p))^2\} \exp\{-\Delta^*_{\Pi,B,q}/g(n,p)\}. \end{aligned} \tag{146}$$

Here $\mathcal{B}_g$ is again a δ-covering set with $\delta = 1/n^2$.

With straightforward calculations, the minimum probability of (145) and (146) goes to zero when $n \to \infty$.

Therefore, we have that

$$\begin{aligned} & L(I,\hat{B},E) - L(\Pi,\hat{B}(\Pi),E) \\ \geq \quad & \Lambda(\mathbf{I},q) - \Lambda(\Pi,q) - O_p(\Delta^*_{\Pi,B,q}). \end{aligned}$$

Finally, noting that $g(n,p) = O(\psi^\sharp_{cb})$ and $V_{\Pi,B,q} \leq V_2(q)$, then it holds that $\Delta_2(X,B^\sharp,\Pi,\Pi^\sharp) \gtrsim x_0$ according to Assumption (39). It then implies $\hat{\Pi} \neq \Pi$ for any Π with $d(\mathbf{I},\Pi) > h_c$ with high probability. This completes the proof.

J. ADMM Computational Approach

For self-completeness, in this section, we discuss the computational aspects of the problem in classical linear models. We relax the ML estimation problem to a bi-convex problem and solve it via an ADMM algorithm proposed in Zhang et al. (2019). A detailed description is given in the sequel.

ADMM formulation First, we are trying to solve

$$\min_{\Pi, B} \|Y - \Pi X B\|_F^2 = \|P_{\Pi X}^{\perp} Y\|_F^2 \tag{147}$$

where projection matrix $P_{\Pi X}^{\perp}$ is defined as $I - \Pi X \left(X^T X\right)^{-1} X^T \Pi^T$. Note that we can decompose Y as $P_{\Pi X}^{\perp} Y + P_{\Pi X} Y$. Since $\|Y\|_F^2 = \|P_{\Pi X}^{\perp} Y\|^2 + \|P_{\Pi X} Y\|^2$ can be treated as a constant, minimizing $\|P_{\Pi X}^{\perp} Y\|^2$ is equivalent to maximizing $\|P_{\Pi X} Y\|^2$.

By introducing two redundant variables Π_1 and Π_2, we formulate (147) as

$$\min_{\Pi_1,\, \Pi_2} -\text{trace}\left(\Pi_1 P_X \Pi_2^T Y Y^T\right), \; s.t. \Pi_1 = \Pi_2, \tag{148}$$

where $P_X := X(X^T X)^{-1} X^T$. We propose to solve (148) with the ADMM Algorithm (Boyd et al., 2011) and present the details of the algorithm in Algorithm 6.

Algorithm 6 ADMM algorithm for the recovery of Π.

1: **Input:** Initial estimate for the permutation matrix $\Pi^{(0)}$ and create an $n \times n$ matrix $\mu^{(0)} = \mathbf{0}$.
2: **For time $t+1$:** Update $\Pi_1^{(t+1)}, \Pi_2^{(t+1)}$ as

$$\begin{aligned}
\Pi_1^{(t+1)} &= \underset{\Pi_1}{\text{argmin}} \langle \Pi_1, -YY^T \Pi_2^{(t)} P_X^T + \mu^{(t)} - \rho \Pi_2^{(t)} \rangle \\
\Pi_2^{(t+1)} &= \underset{\Pi_2}{\text{argmin}} \langle \Pi_2,\; YY^T \Pi_1^{(t+1)} P_X - \mu^{(t)} - \rho \Pi_1^{(t+1)} \rangle \\
\mu^{(t+1)} &= \mu^{(t)} + \rho \left(\Pi_1^{(t+1)} - \Pi_2^{(t+1)}\right).
\end{aligned}$$

3: **Termination**: Stop the ADMM algorithm once $\Pi_1^{(t+1)}$ is identical to $\Pi_2^{(t+1)}$.

Since ADMM may exhibit slow convergence (Boyd et al., 2011), it adopts a warm start strategy to accelerate the algorithm, which consists of two steps:

- Compute the average values $\bar{X} = \frac{1}{p} \sum_{i=1}^{p} X[:, i]$.
- Obtain a rough estimate $\Pi^{(0)}$ by using Algorithm 7 or 8 with $X = \bar{X}$.

Algorithm 7 Averaging estimator.

1: Compute the average $\frac{1}{m} \sum_{i=1}^{m} Y[:, i]$.
2: Compute $\hat{\Pi}$ by maximizing $\left(\langle m^{-1} \sum_{i=1}^{m} Y[:, i], \Pi X \rangle\right)^2$.

Algorithm 8 Eigenvalue estimator.

1: Compute the principal eigenvector $\mathbf{u}$ of $m^{-1} \left(\sum_{i=1}^{m} Y[:, i] Y[:, i]^T\right)$.
2: Recover $\hat{\Pi}$ by maximizing $(\langle \mathbf{u},\ \Pi X \rangle)^2$.