SUPPLEMENT TO "POSTERIOR CONTRACTION RATES OF THE PHYLOGENETIC INDIAN BUFFET PROCESSES"

This manuscript serves as the supplementary material to the paper [9]. Section A and Section B present technical proofs of the main results of the paper. Section C presents an alternative analysis of the real data studied in [9]. Given a matrix $A = (a_{ij})_{m\times n}$, its sup-norm is defined as $||A||_\infty = \max_{ij}|a_{ij}|$ and its spectral norm is defined $||A|| = s_{\max}(A)$, where $s_{\max}(\cdot)$ is the largest singular value of a matrix. The notation $\mathbb{P}$ and $\mathbb{E}$ stand for generic probability and expectation operators when the associated distribution is clear from the context. We use C and its variants such as C' and C_1 to denote generic constants, which may vary from line to line.

A Proofs

A.1 Preparatory lemmas

Lemma A.1. *There is some constant $C > 0$ such that for any $t > 0$,*

$$P_Z\left\{\left\|\frac{1}{p}XX^T - (ZZ^T + I)\right\|_F > t\right\} \le \exp\left\{-Cp\min\left(\frac{t^2}{n^2||ZZ^T+I)||_\infty^2}, \frac{t}{n||ZZ^T+I)||_\infty}\right) + 2\log n\right\},$$

and

$$P_Z\left\{\left\|\frac{1}{p}XX^T - (ZZ^T + I)\right\|_\infty > t\right\} \le \exp\left\{-Cp\min\left(\frac{t^2}{||ZZ^T+I)||_\infty^2}, \frac{t}{||ZZ^T+I)||_\infty}\right) + 2\log n\right\}.$$

A.2 Proofs of Theorem 4.1 and Lemma 4.1

For notational simplicity, we write ϵ for $\epsilon_{n,p}$, with the dependency on n and p being implicit.

Proof of Theorem 4.1. The posterior distribution, according to Bayes formula, is

$$\Pi(U|X) = \frac{\int_U \frac{p(X|Z)}{p(X|Z_0)}d\Pi([Z])}{\int \frac{p(X|Z)}{p(X|Z_0)}d\Pi([Z])}.$$

The denominator has lower bound

$$\int \frac{p(X|Z)}{p(X|Z_0)}d\Pi([Z]) \ge \int_{\{||ZZ^T - Z_0Z_0^T||_F^2=0\}} \frac{p(X|Z)}{p(X|Z_0)}d\Pi([Z]) = \Pi\Big(||ZZ^T - Z_0Z_0^T||_F^2 = 0\Big).$$

The above equality is because $p(X|Z) = p(X|Z_0)$ when $||ZZ^T - Z_0Z_0^T||_F = 0$. Thus, we have

$$\begin{aligned}
E_{Z_0}\Pi(U|X) &\leq E_{Z_0}\phi + E_{Z_0}\Pi(U|X)(1-\phi) \\
&\leq E_{Z_0}\phi + \frac{E_{Z_0}\Big(\int_U \frac{p(X|Z)}{p(X|Z_0)} d\Pi([Z])(1-\phi)\Big)}{\Pi\Big(||ZZ^T - Z_0Z_0^T||_F^2 = 0\Big)} \\
&= E_{Z_0}\phi + \frac{\int_U E_Z(1-\phi) d\Pi([Z])}{\Pi\Big(||ZZ^T - Z_0Z_0^T||_F^2 = 0\Big)} \\
&\leq E_{Z_0}\phi + \frac{1}{\Pi\Big(||ZZ^T - Z_0Z_0^T||_F^2 = 0\Big)} \sup_{Z\in U} E_Z(1-\phi),
\end{aligned}$$

where the equality above is due to Fubini's Theorem. Therefore, the proof is complete. □

Proof of Lemma 4.1. We consider the following test.

$$H_0 : Z = Z_0, \quad H_1 : ||ZZ^T - Z_0Z_0^T||_F > \sqrt{M}\epsilon.$$

The alternative region has decomposition

$$\begin{aligned}
H_1 &\subset \Big\{||ZZ^T - Z_0Z_0^T||_F > \sqrt{M}\epsilon, ||ZZ^T + I||_\infty \leq 4(K_0+1)\Big\} \\
&\quad \cup \bigcup_{l\geq 1} \big\{4l(K_0+1) < ||ZZ^T + I||_\infty \leq 4(l+1)(K_0+1)\big\} \\
&= \bigcup_{l=0}^{\infty} H_{1l}
\end{aligned}$$

Define the testing functions

$$\phi_0 = \mathbb{I}\left\{\left\|\frac{1}{p}XX^T - (Z_0Z_0^T + I)\right\|_F > \frac{1}{2}\sqrt{M}\epsilon\right\},$$

$$\phi_l = \mathbb{I}\left\{\left\|\frac{1}{p}XX^T\right\|_\infty > 2l(K_0+1)\right\}, \quad \text{for each } l.$$

Then, by Lemma A.1 and the fact that $||Z_0Z_0^T + I||_\infty \leq K_0 + 1 \leq 2K_0$, we have

$$E_{Z_0}\phi_0 \leq \exp\left\{-Cp\min\left(\frac{M\epsilon^2}{n^2K_0^2}, \frac{\sqrt{M}\epsilon}{nK_0}\right) + 2\log n\right\},$$

and

$$\begin{aligned}
E_{Z_0}\phi_l &= P_{Z_0}\left\{\left\|\frac{1}{p}XX^T\right\|_\infty > 2l(K_0+1)\right\} \\
&\leq P_{Z_0}\left\{\left\|\frac{1}{p}XX^T - (Z_0Z_0^T + I)\right\|_\infty > 2l(K_0+1) - \left\|Z_0Z_0^T + I\right\|_\infty\right\} \\
&\leq P_{Z_0}\left\{\left\|\frac{1}{p}XX^T - (Z_0Z_0^T + I)\right\|_\infty > l(K_0+1)\right\} \\
&\leq \exp\Big(-Clp + 2\log n\Big),
\end{aligned}$$

where the second inequality above is by $\left\|Z_0Z_0^T+I\right\|_\infty \leq K_0+1 \leq l(K_0+1)$, and the last inequality above is by Lemma A.1. We also have for any $Z \in H_{10}$,

$$\begin{aligned}
E_Z(1-\phi_0) &= P_Z\left\{\left\|\frac{1}{p}XX^T-(Z_0Z_0^T+I)\right\|_F \leq \frac{1}{2}\sqrt{M}\epsilon\right\} \\
&\leq P_Z\left\{||ZZ^T-Z_0Z_0^T||_F-\left\|\frac{1}{p}XX^T-(ZZ^T+I)\right\|_F \leq \frac{1}{2}\sqrt{M}\epsilon\right\} \\
&\leq P_Z\left\{\left\|\frac{1}{p}XX^T-(ZZ^T+I)\right\|_F > \frac{1}{2}\sqrt{M}\epsilon\right\} \\
&\leq \exp\left\{-Cp\min\left(\frac{M\epsilon^2}{n^2K_0^2},\frac{\sqrt{M}\epsilon}{nK_0}\right)+2\log n\right\},
\end{aligned}$$

where the last inequality is by Lemma A.1 and the fact that $||ZZ^T+I||_\infty \leq 4(K_0+1) \leq 8K_0$ for any $Z \in H_{10}$. Taking supreme over $Z \in H_{10}$, we get

$$\sup_{Z\in H_{10}} E_Z(1-\phi_0) \leq \exp\left\{-Cp\min\left(\frac{M\epsilon^2}{n^2K_0^2},\frac{\sqrt{M}\epsilon}{nK_0}\right)+2\log n\right\}.$$

For any $Z \in H_{1l}$, we have

$$\begin{aligned}
E_Z(1-\phi_l) &= P_Z\left\{\left\|\frac{1}{p}XX^T\right\|_\infty \leq 2l(K_0+1)\right\} \\
&\leq P_Z\left\{||ZZ^T+I||_\infty-\left\|\frac{1}{p}XX^T-(ZZ^T+I)\right\|_\infty \leq 2l(K_0+1)\right\} \\
&\leq P_Z\left\{\left\|\frac{1}{p}XX^T-(ZZ^T+I)\right\|_\infty > 2l(K_0+1)\right\} \\
&\leq \exp\Big(-Cp+2\log n\Big),
\end{aligned}$$

where the last inequality above uses Lemma A.1 and the fact that $||ZZ^T+I||_\infty \leq 4(l+1)(K_0+1)$ for $Z \in H_{1l}$, and the second last inequality uses the fact that $||ZZ^T+I||_\infty > 4l(K_0+1)$ for all $Z \in H_{1l}$. Taking supreme over $Z \in H_{1l}$, we obtain

$$\sup_{Z\in H_{1l}} E_Z(1-\phi_l) \leq \exp\Big(-Cp+2\log n\Big).$$

Define $\phi = \max_l \phi_l$, we have

$$\begin{aligned} & E_{Z_0}\phi + \sup_{Z \in H_1} E_Z(1-\phi) \\ = \quad & E_{Z_0} \max_l \phi_l + \max_l \sup_{Z \in H_{1l}} E_Z(1-\phi) \\ \leq \quad & \sum_l E_{Z_0}\phi_l + \max_l \sup_{Z \in H_{1l}} E_Z(1-\phi_l) \\ \leq \quad & 2\exp\left\{-Cp\min\left(\frac{M\epsilon^2}{n^2K_0^2}, \frac{\sqrt{M}\epsilon}{nK_0}\right) + 2\log n\right\} \\ & + \sum_{l=1}^{\infty} \exp\Big(-Clp + 2\log n\Big) + \exp\left(-Cp + 2\log n\right) \\ \leq \quad & 2\exp\left\{-Cp\min\left(\frac{M\epsilon^2}{n^2K_0^2}, \frac{\sqrt{M}\epsilon}{nK_0}\right) + 2\log n\right\} + \exp\Big(-C'p + 2\log n\Big). \end{aligned}$$

Thus, the proof is complete. □

A.3 Proof of Theorems 4.2-4.4

Proof of Theorem 4.2. Without loss of generality, we assume n is even in the proof. First, note that the event $\{||ZZ^T - Z_0Z_0^T||_F^2 = 0\}$ is implied by $\{||Z - Z_0||_F^2 = 0\}$ for any column ordering of Z_0. Therefore, we have

$$\Pi\Big(||ZZ^T - Z_0Z_0^T||_F^2 = 0\Big) \geq P\Big(||Z - Z_0||_F^2 = 0\Big),$$

with P being any probability measure on Z whose image measure under the map $Z \mapsto [Z]$ is pIBP. We choose P to be the stick-breaking representation described in Section 3.2. That is, under probability P, we first sample $\{p_k\}$ according to (2) in [9], and then given $\{p_k\}$, Z is sampled according to the two-group tree structure for each column. Define r_{1k} and r_{2k} to be the group nodes for the first and the second group, respectively, for each k. Then according to the stick-breaking representation of pIBP, $\{r_{1k}\}$ and $\{r_{2k}\}$ given $\{p_k\}$ are i.i.d. Bernoulli random variables with parameter $1 - \exp(-\eta\gamma_k)$, where $\gamma_k = -\log(1 - p_k)$. Then, z_{ik} are sampled conditioning on (r_{1k}, r_{2k}). When $r_{1k} = 1$, we have $z_{ik} = 1$ for all $i \in S_1$. When $r_{1k} = 0$, z_{ik} follows the Bernoulli distribution with parameter $1 - \exp\big(-(1-\eta)\gamma_k\big)$ for all $i \in S_1$. The value of r_{2k} determines the distribution of z_{ik} for $i \in S_2$ in the same way.

We first study $P\Big(||Z - Z_0||_F^2 = 0\Big|\{v_k\}, \alpha\Big)$ for given $\{v_k\}$ and α. We choose a particular ordering of columns of Z_0. Given the factor decomposition (5) in [9], let the first K_0^* columns correspond to the group-shared factors, and the next $K_{01} + K_{02}$ columns correspond to the group specific factors. Then define the number of 1's in the k-th column of Z_0 by

$$m_k = \sum_{\{i: z_{0,ik}=1\}} z_{0,ik}, \quad \text{for } k = 1, ..., K_0^*.$$

Define $M^* = \sum_{k=1}^{K_0^*} m_k$ to be the number of 1's in the first K_0^* columns of Z_0. The quantity $||Z - Z_0||_F^2$ has four parts.

$$||Z - Z_0||_F^2 = \sum_{k=1}^{K_0^*} U_k + \sum_{k=1}^{K_0^*} V_k + \sum_{k=K_0^*+1}^{K_0} \sum_{i=1}^{n} (z_{ik} - z_{0,ik})^2 + \sum_{k=K_0+1}^{\infty} \sum_{i=1}^{n} z_{ik}.$$

where

$$U_k = \sum_{\{i: z_{0,ik}=0\}} z_{ik}, \quad V_k = \sum_{\{i: z_{0,ik}=1\}} |z_{ik} - 1|.$$

We observe that given $\{v_k\}$, the four terms are independent. Therefore

$$\begin{aligned}
& P\Big(||Z - Z_0||_F^2 = 0 | \{v_k\}, \alpha\Big) \\
= \; & P\left(\sum_{k=1}^{K_0^*} U_k = 0 \Big| \{v_k\}, \alpha\right) \times P\left(\sum_{k=1}^{K_0^*} V_k = 0 \Big| \{v_k\}, \alpha\right) \\
& \times P\left(\sum_{k=K_0^*+1}^{K_0} \sum_{i=1}^{n} (z_{ik} - z_{0,ik})^2 = 0 \Big| \{v_k\}, \alpha\right) \times P\left(\sum_{k=K_0+1}^{\infty} \sum_{i=1}^{n} z_{ik} = 0 \Big| \{v_k\}, \alpha\right). \quad (8)
\end{aligned}$$

We study the four terms separately. Define $\mathcal{H} = \left\{\frac{1}{4} \leq v_i \leq \frac{3}{4}, \text{ for } k = 1, ..., K_0\right\}$. Then, for every $\{v_k\} \in \mathcal{H}$, we have

$$\begin{aligned}
& P\left(\sum_{k=1}^{K_0^*} U_k = 0 \Big| \{v_k\}, \alpha\right) \times P\left(\sum_{k=1}^{K_0^*} V_k = 0 \Big| \{v_k\}, \alpha\right) \\
\geq \; & \Big(\exp\big(-\gamma_1(1-\eta)\big)\Big)^{nK_0^* - M^*} \Big(1 - \exp\big(-\gamma_{K_0^*}(1-\eta)\big)\Big)^{M^*} \\
& \times P\Big(r_{11} = ... = r_{1K_0^*} = r_{21} = ... = r_{2K_0^*} = 0 \Big| \{v_k\}, \alpha\Big) \\
\geq \; & \Big(\exp\big(-\gamma_1(1-\eta)\big)\Big)^{nK_0^* - M^*} \Big(1 - \exp\big(-\gamma_{K_0^*}(1-\eta)\big)\Big)^{M^*} \times \exp\Big(-2K_0^*\gamma_1\eta\Big) \\
= \; & (1-p_1)^{(nK_0^* - M^*)(1-\eta) + 2K_0^*\eta} \left(1 - (1 - p_{K_0^*})^{1-\eta}\right)^{M^*} \\
\geq \; & (1-p_1)^{(nK_0^* - M^*)(1-\eta) + 2K_0^*\eta} p_{K_0^*}^{M^*} (1-\eta)^{M^*} \quad (9) \\
\geq \; & 4^{-(nK_0^* - M^*)(1-\eta)} 4^{-2K_0^*\eta} 4^{-K_0^* M^*} (1-\eta)^{M^*} \\
\geq \; & \exp(-CnK_0^{*2})(1-\eta)^{nK_0^*}, \quad (10)
\end{aligned}$$

where we have used the inequality $1 - q^\beta \geq \beta(1-q)$ for $\beta, q \in (0,1)$ to derive (9). The inequality (10) is due to the bound $M^* \le nK_0^*$. The third term of (8) is

$$
\begin{aligned}
& P\Big(\sum_{k=K_0^*+1}^{K_0}\sum_{i=1}^{n}(z_{ik}-z_{0,ik})^2=0\Big|\{v_k\},\alpha\Big)\\
\ge\;& \exp\Big(-n(K_{01}+K_{02})\gamma_{K_0^*}(1-\eta)/2\Big)\times P\Big(r_{1k}=1,r_{2k}=0,\ \text{for } k=K_0^*+1,...,K_0^*+K_{01}\Big|\{v_k\},\alpha\Big)\\
& \times P\Big(r_{1k}=0,r_{2k}=1,\ \text{for } k=K_0^*+K_{01}+1,...,K_0^*+K_{01}+K_{02}\Big|\{v_k\},\alpha\Big)\\
\ge\;& \exp\Big(-n(K_{01}+K_{02})\gamma_{K_0^*}(1-\eta)/2\Big)\times\Big(1-\exp(-\eta\gamma_{K_0})\Big)^{K_{01}+K_{02}}\times\exp\Big(-\eta(K_{01}+K_{02})\gamma_{K_0^*}\Big)\\
\ge\;& (1-p_{K_0^*})^{(\eta+n(1-\eta)/2)(K_{01}+K_{02})}p_{K_0}^{K_{01}+K_{02}}\eta^{K_{01}+K_{02}}\\
\ge\;& \Big(1-(4/3)^{-K_0^*}\Big)^{(\eta+n(1-\eta)/2)(K_{01}+K_{02})}4^{-K_0(K_0-K_0^*)}\eta^{K_{01}+K_{02}}\\
=\;& \exp\Big((\eta+n(1-\eta)/2)(K_{01}+K_{02})\log\Big(1-(4/3)^{-K_0^*}\Big)\Big)4^{-K_0(K_0-K_0^*)}\eta^{K_{01}+K_{02}}\\
\ge\;& \exp\Big(-Cn\frac{K_0-K_0^*}{(4/3)^{K_0^*}}-CK_0(K_0-K_0^*)\Big)\eta^{K_{01}+K_{02}},
\end{aligned}
$$

where the last inequality is due to the fact that $\log(1-x)\ge-\delta x, \quad \text{for } |x|\le 3/4$, with $\delta>0$ being a universal constant. The last term in the product (8) is

$$
\begin{aligned}
& P\Big(\sum_{k=K_0+1}^{\infty}\sum_{i=1}^{n}z_{ik}=0\Big|\{v_k\},\alpha\Big)\\
\ge\;& \prod_{k=K_0+1}^{\infty}\exp\Big(-n\gamma_k(1-\eta)\Big)\times P\Big(r_{1k}=r_{2k}=0,\ \text{for } k>K_0\Big|\{v_k\},\alpha\Big)\\
\ge\;& \prod_{k=K_0+1}^{\infty}\exp\Big(-n\gamma_k(1-\eta)\Big)\times\prod_{k=K_0+1}^{\infty}\exp(-2\eta\gamma_k)\\
=\;& \prod_{k=K_0+1}^{\infty}(1-p_k)^{n(1-\eta)+2\eta}\\
\ge\;& \prod_{k=K_0+1}^{\infty}\Big(1-(4/3)^{-k}\Big)^{n(1-\eta)+2\eta}\\
=\;& \exp\left((n(1-\eta)+2\eta)\sum_{k=K_0+1}^{\infty}\log\Big(1-(4/3)^{-k}\Big)\right)\\
\ge\;& \exp\left(-\delta(n(1-\eta)+2\eta)\sum_{k=K_0+1}^{\infty}(4/3)^{-k}\right) \qquad (11)\\
=\;& \exp\Big(-3\delta(4/3)^{-K_0}(n(1-\eta)+2\eta)\Big)\\
\ge\;& \exp(-Cn)
\end{aligned}
$$

where the inequality (11) is due to the fact that $\log(1-x)\ge-\delta x, \quad \text{for } |x|\le 3/4$, with $\delta>0$ being a universal constant. For a constant $\eta\in(0,1)$, we have $(1-\eta)^{nK_0^*}\ge\exp(-CnK_0^*)$

and $\eta^{K_{01}+K_{02}} \geq \exp(-C(K_0 - K_0^*))$ and thus

$$
\begin{aligned}
& P\Big(||Z - Z_0||_F^2 = 0|\{v_k\}, \alpha\Big) \\
\geq \;& \exp\left(-Cn(K_0^{*2}+1) - Cn\frac{K_0 - K_0^*}{(4/3)^{K_0^*}} - CK_0(K_0 - K_0^*)\right),
\end{aligned}
$$

for every $\{v_k\} \in \mathcal{H}$. Observe that the above argument also works by replacing K_0 and K_0^* by $K_0 + \kappa$ and $K_0^* + \kappa$ for any $\kappa \geq 0$. Thus, for a constant $\eta \in (0, 1)$, we have

$$
\begin{aligned}
& P\Big(||Z - Z_0||_F^2 = 0|\{v_k\}, \alpha\Big) \\
\geq \;& \exp\left(-Cn((K_0^*+\kappa)^2+1) - Cn\frac{K_0 - K_0^*}{(4/3)^{K_0^*+\kappa}} - C(K_0+\kappa)(K_0 - K_0^*)\right), \qquad (12)
\end{aligned}
$$

for every $\{v_k\} \in \mathcal{H}$. When $\eta = 0$, pIBP becomes IBP. Thus, the decomposition (8) becomes

$$
\begin{aligned}
& P\Big(||Z - Z_0||_F^2 = 0|\{v_k\}, \alpha\Big) \\
= \;& P\left(\sum_{k=1}^{K_0} U_k = 0\Big|\{v_k\}, \alpha\right) \times P\left(\sum_{k=1}^{K_0} V_k = 0\Big|\{v_k\}, \alpha\right) \times P\left(\sum_{k=K_0+1}^{\infty}\sum_{i=1}^{n} z_{ik} = 0\Big|\{v_k\}, \alpha\right).
\end{aligned}
$$

Replacing K_0^* by K_0 in (10), we have

$$
P\Big(||Z - Z_0||_F^2 = 0|\{v_k\}, \alpha\Big) \geq \exp\left(-CnK_0^2\right),
$$

for $\eta = 0$ and every $\{v_k\} \in \mathcal{H}$. Note that this is a special case of (12) with $K_0^* = K_0$ and $\kappa = 0$. Finally, we have

$$
\begin{aligned}
& P\Big(||Z - Z_0||_F^2 = 0\Big) \\
\geq \;& P\Big(||Z - Z_0||_F^2 = 0\Big|\{v_k\} \in \mathcal{H}, \alpha \in (1/2, 2)\Big)\mathbb{P}\Big(\mathcal{H}\Big|\alpha \in (1/2, 2)\Big)\mathbb{P}\Big(\alpha \in (1/2, 2)\Big) \\
\geq \;& P\Big(||Z - Z_0||_F^2 = 0\Big|\{v_k\} \in \mathcal{H}, \alpha \in (1/2, 2)\Big)\left(\sup_{\alpha \in (1/2,2)} \frac{\alpha \mathrm{B}(\alpha, 1)}{(3/4)^\alpha - (1/4)^\alpha}\right)^{-K_0-\kappa}\mathbb{P}\Big(\alpha \in (1/2, 2)\Big) \\
\geq \;& \exp(-C(K_0+\kappa))P\Big(||Z - Z_0||_F^2 = 0\Big|\{v_k\} \in \mathcal{H}, \alpha \in (1/2, 2)\Big) \\
\geq \;& \exp\left(-Cn((K_0^*+\kappa)^2+1) - Cn\frac{K_0 - K_0^*}{(4/3)^{K_0^*+\kappa}} - C(K_0+\kappa)(K_0 - K_0^*+1)\right),
\end{aligned}
$$

by plugging (12). Thus, the proof is complete. □

Proof of Theorem 4.3-4.4. This is directly by combining Theorem 4.1, Lemma 4.1, Theorem 4.2 and the discussion after Theorem 4.2. For Theorem 4.3, we have

$$
\begin{aligned}
& E_{Z_0}\Pi\left(\left\|ZZ^T - Z_0Z_0^T\right\|^2 > M\epsilon^2|X\right) \\
\leq \;& \frac{\exp\left\{-Cp\min\left(\frac{M\epsilon^2}{n^2K_0^2}, \frac{\sqrt{M}\epsilon}{nK_0}\right) + 2\log n\right\} + \exp\Big(-Cp + 2\log n\Big)}{\exp\left(-C_1nK_0^2\right)}.
\end{aligned}
$$

Taking $\epsilon^2 = \frac{K_0^4 n^3}{p}$, we have $p \min\left(\frac{M\epsilon^2}{n^2 K_0^2}, \frac{\sqrt{M}\epsilon}{nK_0}\right) \asymp nK_0^2$ under the assumption of Theorem 4.3 that $nK_0^2 = o(p)$. Thus, for some sufficiently large M,

$$E_{Z_0}\Pi\left(\left\|ZZ^T - Z_0 Z_0^T\right\|^2 > M\epsilon^2 | X\right) \leq \exp\left(-C' nK_0^2\right) + \exp\left(-C'p + 2\log n\right).$$

Note that the first term in the tail dominates, which gives the result of Theorem 4.3. The result of Theorem 4.4 follows a similar argument. □

A.4 Unknown Variances

When variances $(\sigma_{A,0}^2, \sigma_{X,0}^2)$ are unknown, we put independent prior $\pi = \pi_A \times \pi_X$ on them, so that

$$([Z], \sigma_A^2, \sigma_X^2) \sim \Pi = \pi_{[Z]} \times \pi_A \times \pi_X,$$

where $\pi_{[Z]}$ is pIBP or IBP on $[Z]$. In this case, we use the following theorem instead of Theorem 4.1.

Theorem A.1. *Assume*

$$\Pi\Big((2\sigma_X^4)^{-1}||\sigma_A^2 ZZ^T + \sigma_X^2 I - (\sigma_{A,0}^2 Z_0 Z_0^T + \sigma_{X,0}^2 I)||_F^2 \leq \epsilon^2\Big) \geq \exp\Big(-Cp\epsilon^2\Big), \tag{13}$$

for some ϵ satisfying $p\epsilon^2 \to \infty$ and some constant $C > 0$, and there is a testing function ϕ, such that $E_{Z_0}\phi + \sup_{Z\in U} E_Z(1-\phi) \leq \exp\Big(-(C+4)p\epsilon^2\Big)$, then

$$E_{Z_0}\Pi\Big(U|X\Big) \leq \frac{C'}{p\epsilon^2},$$

for some constant $C' > 0$.

Proof. In view of Theorem 2.1 of [17], we only need to lower bound the prior probability of the Kullback-Leibler neighborhood of the truth. That is, we need to show that (13) implies

$$\Pi\left\{E_{Z_0}\left(\log\frac{dP_{Z_0}}{dP_Z}\right) \vee \mathrm{Var}_{P_{Z_0}}\left(\log\frac{dP_{Z_0}}{dP_Z}\right) \leq \epsilon^2\right\} \geq \exp\Big(-Cp\epsilon^2\Big).$$

According to (1) in [9], we have

$$P_Z = N(0, \Sigma) \quad \text{and} \quad P_{Z_0} = N(0, \Sigma_0),$$

where we use the notation $\Sigma = \sigma_A^2 ZZ^T + \sigma_X^2 I$ and $\Sigma_0 = \sigma_{A,0}^2 Z_0 Z_0^T + \sigma_{X,0}^2 I$. The same proof of Lemma 8 in [16] can be applied to derive the bounds

$$E_{Z_0}\left(\log\frac{dP_{Z_0}}{dP_Z}\right) \vee \mathrm{Var}_{P_{Z_0}}\left(\log\frac{dP_{Z_0}}{dP_Z}\right) \leq \frac{1}{2}\left\|(\Sigma - \Sigma_0)\Sigma^{-1}\right\|_F^2. \tag{14}$$

We bound $\frac{1}{2}\left\|(\Sigma-\Sigma_0)\Sigma^{-1}\right\|_F^2$ by

$$\begin{aligned}
& \frac{1}{2}\left\|\Sigma-\Sigma_0\right\|_F^2 ||\Sigma^{-1}||^2 \\
= \; & \frac{1}{2}\left\|\Big(\sigma_A^2 ZZ^T+\sigma_X^2 I-(\sigma_{A,0}^2 Z_0 Z_0^T+\sigma_{X,0}^2 I)\Big)\right\|_F^2 \left\|\Big(\sigma_A^2 ZZ^T+\sigma_X^2 I\Big)^{-1}\right\|^2 \\
\leq \; & \frac{1}{2\sigma_X^4}\left\|\sigma_A^2 ZZ^T+\sigma_X^2 I-(\sigma_{A,0}^2 Z_0 Z_0^T+\sigma_{X,0}^2 I)\right\|_F^2,
\end{aligned}$$

where the last inequality is because

$$\left\|\Big(\sigma_A^2 ZZ^T+\sigma_X^2 I\Big)^{-1}\right\| \leq \Big(\lambda_{\min}\Big(\sigma_A^2 ZZ^T+\sigma_X^2 I\Big)\Big)^{-1} \leq \sigma_X^{-2}.$$

Therefore, we have

$$\begin{aligned}
& \Pi\left\{E_{Z_0}\left(\log\frac{dP_{Z_0}}{dP}\right)\vee \mathrm{Var}_{P_{Z_0}}\left(\log\frac{dP_{Z_0}}{dP}\right)\leq \epsilon^2\right\} \\
\geq \; & \Pi\left\{\frac{1}{2\sigma_X^4}\left\|\sigma_A^2 ZZ^T+\sigma_X^2 I-(\sigma_{A,0}^2 Z_0 Z_0^T+\sigma_{X,0}^2 I)\right\|_F^2 \leq \epsilon^2\right\} \\
\geq \; & \exp\Big(-Cp\epsilon^2\Big).
\end{aligned}$$

Thus, the proof is complete. □

Theorem A.2. *Assume* $\log p \lesssim n$. *Theorem 4.3 and 4.4 still hold if there are universal constants* $B>0$ *and* $C>0$, *such that* $\sigma_{A,0}^2 \in (B^{-1}, B)$, $\sigma_{X,0}^2 \in (B^{-1}, B)$ *and* $\inf_{t\in(0,2B)}\pi_A(t)\wedge \inf_{t\in(0,2B)}\pi_X(t)\geq CB^{-1}$.

Proof. According to Theorem A.1 and Lemma 4.1, we only need to show

$$\log\Pi\Big((2\sigma_X^4)^{-1}||\sigma_A^2 ZZ^T+\sigma_X^2 I-(\sigma_{A,0}^2 Z_0 Z_0^T+\sigma_{X,0}^2 I)||_F^2\leq\epsilon^2\Big)$$

can be lower bounded by the same order of prior mass in all situations considered in Section 4.4. Using conditioning and the independent structure of the prior, we have

$$\begin{aligned}
& \Pi\Big((2\sigma_X^4)^{-1}||\sigma_A^2 ZZ^T+\sigma_X^2 I-(\sigma_{A,0}^2 Z_0 Z_0^T+\sigma_{X,0}^2 I)||_F^2\leq\epsilon^2\Big) \\
\geq \; & \Pi\Big((2\sigma_X^4)^{-1}||(\sigma_{A,0}^2-\sigma_A^2)Z_0Z_0^T+(\sigma_{X,0}^2-\sigma_X^2)I||_F^2\leq\epsilon^2\Big)\Pi\Big(||ZZ^T-Z_0Z_0^T||_F^2=0\Big) \\
\geq \; & \Pi\left(n^2K_0\left|\frac{\sigma_{A,0}^2-\sigma_A^2}{\sigma_X^2}\right|^2+n\left|\frac{\sigma_{X,0}^2-\sigma_X^2}{\sigma_X^2}\right|^2\leq\epsilon^2\right)\Pi\Big(||ZZ^T-Z_0Z_0^T||_F^2=0\Big),
\end{aligned}$$

because $||Z_0Z_0^T||_F^2 \le n^2K_0$ and $||I||_F^2 = n$. The variance part has lower bound

$$\begin{aligned}
&\Pi\left(n^2K_0\left|\frac{\sigma_{A,0}^2-\sigma_A^2}{\sigma_X^2}\right|^2+n\left|\frac{\sigma_{X,0}^2-\sigma_X^2}{\sigma_X^2}\right|^2\le\epsilon^2\right)\\
\ge\ &\Pi\left(n^2K_0\left|\frac{\sigma_{A,0}^2-\sigma_A^2}{\sigma_X^2}\right|^2\le\epsilon^2/2, n\left|\frac{\sigma_{X,0}^2-\sigma_X^2}{\sigma_X^2}\right|^2\le\epsilon^2/2\right)\\
\ge\ &\Pi\left(n^2K_0B^2\Big(1+\epsilon/\sqrt{2n}\Big)^2|\sigma_A^2-\sigma_{A,0}^2|^2\le\epsilon^2/2, n\left|\frac{\sigma_{X,0}^2-\sigma_X^2}{\sigma_X^2}\right|^2\le\epsilon^2/2\right)\\
=\ &\pi_A\left(n^2K_0B^2\Big(1+\epsilon/\sqrt{2n}\Big)^2|\sigma_A^2-\sigma_{A,0}^2|^2\le\epsilon^2/2\right)\pi_X\left(n\left|\frac{\sigma_{X,0}^2-\sigma_X^2}{\sigma_X^2}\right|^2\le\epsilon^2/2\right).
\end{aligned}$$

We give lower bounds for the two terms above separately. When $\frac{\epsilon^2}{2n}$ does not go to 0, $\pi_X\left(n\left|\frac{\sigma_{X,0}^2-\sigma_X^2}{\sigma_X^2}\right|^2\le\epsilon^2/2\right)$ can be lower bounded by a constant. When it goes to 0, we have

$$\begin{aligned}
\pi_X\left(n\left|\frac{\sigma_{X,0}^2-\sigma_X^2}{\sigma_X^2}\right|^2\le\epsilon^2/2\right)\ &\ge\ \int_{\frac{\sigma_{X,0}^2\sqrt{2n}}{\sqrt{2n}+\epsilon}}^{\frac{\sigma_{X,0}^2\sqrt{2n}}{\sqrt{2n}-\epsilon}}\pi_X(t)dt\\
&\ge\ C_1B^{-2}\frac{\epsilon}{\sqrt{n}}.
\end{aligned}$$

Similarly, when $\frac{\epsilon^2}{(1+\epsilon/\sqrt{2n})^2}$ does not go to 0, $\pi_A\left(n^2K_0B^2\Big(1+\epsilon/\sqrt{2n}\Big)^2|\sigma_A^2-\sigma_{A,0}^2|^2\le\epsilon^2/2\right)$ can be lower bounded by a constant. When it goes to zero, we have

$$\begin{aligned}
&\pi_A\left(n^2K_0B^2\Big(1+\epsilon/\sqrt{2n}\Big)^2|\sigma_A^2-\sigma_{A,0}^2|^2\le\epsilon^2/2\right)\\
\ge\ &\frac{C_2\epsilon}{n\sqrt{K_0}B^2\Big(1+\epsilon/\sqrt{n}\Big)}.
\end{aligned}$$

To summarize, for any rate ϵ appearing in Theorems 4.3 and 4.4, we have

$$\Pi\left(n^2K_0\left|\frac{\sigma_{A,0}^2-\sigma_A^2}{\sigma_X^2}\right|^2+n\left|\frac{\sigma_{X,0}^2-\sigma_X^2}{\sigma_X^2}\right|^2\le\epsilon^2\right)\ge\exp\Big(-C'\big(\log p+\log n+\log K_0\big)\Big),$$

for a constant C_0 only depending on B. Hence, for Theorem 4.3, we have

$$\begin{aligned}
&\Pi\Big((2\sigma_X^4)^{-1}||\sigma_A^2ZZ^T+\sigma_X^2I-(\sigma_{A,0}^2Z_0Z_0^T+\sigma_{X,0}^2I)||_F^2\le\epsilon^2\Big)\\
\ge\ &\exp\Big(-C'\big(\log p+\log n+\log K_0\big)\Big)\times\exp\Big(-CnK_0^2\Big)\\
\ge\ &\exp\Big(-C_1nK_0^2\Big),
\end{aligned}$$

for some $C_1 > 0$ because $\log p \lesssim n$. Combining this lower bound with Lemma 4.1, the conditions of Theorem A.1 holds for $\epsilon^2 = nK_0^2/p$ and

$$U = \left\{ ||ZZ^T - Z_0 Z_0^T||_F^2 > M \frac{K_0^4 n^3}{p} \right\},$$

which implies $E_{Z_0}\Pi(U|X) \to 0$. For Theorem 4.4, we have

$$\begin{aligned} & \Pi\Big((2\sigma_X^4)^{-1}||\sigma_A^2 ZZ^T + \sigma_X^2 I - (\sigma_{A,0}^2 Z_0 Z_0^T + \sigma_{X,0}^2 I)||_F^2 \le \epsilon^2\Big) \\ \ge\ & \exp\Big(- C'\big(\log p + \log n + \log K_0\big)\Big) \times \exp\Big(- CnK_0^{2(1-\beta)}\Big) \\ \ge\ & \exp\Big(- C_2 n K_0^{2(1-\beta)}\Big), \end{aligned}$$

for some $C_2 > 0$. Combining this lower bound with Lemma 4.1, the conditions of Theorem A.1 holds for $\epsilon^2 = nK_0^{2(1-\beta)}/p$ and

$$U = \left\{ ||ZZ^T - Z_0 Z_0^T||_F^2 > M \frac{K_0^{4-2\beta} n^3}{p} \right\},$$

which implies $E_{Z_0}\Pi(U|X) \to 0$. □

A.5 Misspecified Structure

To handle misspecified structure, we need an argument involving a change of measure. The following bound is a general result for all prior distributions Π.

Lemma A.2. *For any $Z_0 \in \{0,1\}^{n\times K_0}$ and $Z^* \in \{0,1\}^{n\times K^*}$, the following inequality holds for any measurable set U,*

$$E_{Z_0}\Pi(U|X) \le \exp\left(p\|Z_0 Z_0^T - Z^*(Z^*)^T\|_{\mathrm{F}}^2\right) E_{Z^*}\Pi(U|X) + \frac{2}{p\|Z_0 Z_0^T - Z^*(Z^*)^T\|_{\mathrm{F}}^2}.$$

Proof. Let us use the notation

$$P_{Z_0} = N(0, Z_0 Z_0^T + I), \quad \text{and} \quad P_{Z^*} = N(0, Z^*(Z^*)^T + I).$$

By (14) and the bound $\left\|(Z^*(Z^*)^T + I)^{-1}\right\| \le 1$, we have

$$E_{Z_0}\left(\sum_{j=1}^p \log \frac{dP_{Z_0}}{dP_{Z^*}}(X_j)\right) \vee \mathrm{Var}_{Z_0}\left(\sum_{j=1}^p \log \frac{dP_{Z_0}}{dP_{Z^*}}(X_j)\right) \le \frac{1}{2} p\|Z_0 Z_0^T - Z^*(Z^*)^T\|_{\mathrm{F}}^2.$$

Define the event

$$B = \left\{ \frac{p(X|Z_0)}{p(X|Z^*)} \le \exp\left(p\|Z_0 Z_0^T - Z^*(Z^*)^T\|_{\mathrm{F}}^2\right) \right\}.$$

By Chebyshev's inequality,

$$\begin{aligned}
P_{Z_0}(B^c) &= P_{Z_0}\left\{\log\frac{p(X|Z_0)}{p(X|Z^*)} > p\|Z_0Z_0^T - Z^*(Z^*)^T\|_{\mathrm{F}}^2\right\} \\
&\leq P_{Z_0}\left\{\sum_{j=1}^{p}\left(\log\frac{dP_{Z_0}}{dP_{Z^*}}(X_j) - E_{Z_0}\left(\frac{dP_{Z_0}}{dP_{Z^*}}\right)\right) > \frac{1}{2}p\|Z_0Z_0^T - Z^*(Z^*)^T\|_{\mathrm{F}}^2\right\} \\
&\leq \frac{2}{p\|Z_0Z_0^T - Z^*(Z^*)^T\|_{\mathrm{F}}^2}.
\end{aligned}$$

Therefore, for any U,

$$\begin{aligned}
E_{Z_0}\Pi(U|X) &\leq E_{Z_0}\Pi(U|X)\mathbb{I}_B + P_{Z_0}(B^c) \\
&= E_{Z^*}\frac{p(X|Z_0)}{p(X|Z^*)}\Pi(U|X)\mathbb{I}_B + P_{Z_0}(B^c) \\
&\leq \exp\left(p\|Z_0Z_0^T - Z^*(Z^*)^T\|_{\mathrm{F}}^2\right)E_{Z^*}\Pi(U|X) + \frac{2}{p\|Z_0Z_0^T - Z^*(Z^*)^T\|_{\mathrm{F}}^2}.
\end{aligned}$$

The proof is complete. □

To apply this result, let us consider a binary factor matrix $Z_0 \in \{0,1\}^{n\times K_0}$. It is close to a binary matrix $Z^* \in \{0,1\}^{n\times K_0}$ which has a well-specified group structure with $K_0^* \lesssim K_0^{1-\beta}$. Then, Lemma A.2 allows one to bound the posterior probability under the true model $E_{Z_0}\Pi(U|X)$ by $E_{Z^*}\Pi(U|X)$. The object $E_{Z^*}\Pi(U|X)$ can be well bounded because Z^* has an exact two-group structure.

To make this idea work, we need a strengthened version of Theorem 4.4 in the paper with a faster tail probability for certain technical reasons. This can be achieved by the following two lemmas.

Lemma A.3. *For an arbitrary $Z_0 \in \{0,1\}^{n\times K_0}$, under the assumption of Theorem 4.4, there exist some constants $C_1, C_2 > 0$, such that*

$$E_{Z_0}\Pi\left(\left\|ZZ^T + I\right\|_\infty \leq C_1(K_0+1)|X\right) \geq 1 - \exp\left(-C_2 p\right).$$

Proof. We prove the result using the general inequality established in Theorem 4.1 for $U = \{\left\|ZZ^T + I\right\|_\infty > C_1K_0\}$. In view of the prior mass lower bound in Theorem 4.2, it is sufficient to establish a test with desired error probability for

$$H_0 : Z = Z_0, \quad H_1 : \left\|ZZ^T + I\right\|_\infty > C_1(K_0+1).$$

Let us decompose the alternative set by

$$H_1 \subset \bigcup_{l\geq 1}\left\{C_1 l(K_0+1) < ||ZZ^T + I||_\infty \leq C_1(l+1)(K_0+1)\right\} = \bigcup_{l\geq 1} H_{1l}.$$

Following the proof of Lemma 4.1, there exists ϕ_l for each $l \geq 1$, such that

$$E_{Z_0}\phi_l \leq \exp\left(-Clp + 2\log n\right),$$

and

$$\sup_{Z\in H_{1l}} E_Z(1-\phi_l) \le \exp\left(-Cp + 2\log n\right).$$

Define $\phi = \max_{l\ge 1}\phi_l$, and then we have

$$\begin{aligned} E_{Z_0}\phi + \sup_{Z\in H_1} E_Z(1-\phi) &\le E_{Z_0}\phi + \max_{l\ge 1}\sup_{Z\in H_{1l}} E_Z(1-\phi) \\ &\le \sum_{l\ge 1} E_{Z_0}\phi_l + \max_{l\ge 1}\sup_{Z\in H_{1l}} E_Z(1-\phi_l) \\ &\le \sum_{l\ge 1}\exp\left(-Clp + 2\log n\right) + \exp\left(-Cp + 2\log n\right) \\ &\le 2\exp\left(-C'p + 2\log n\right). \end{aligned}$$

The result follows by applying Theorem 4.1 and the prior mass lower bound in Theorem 4.2. □

Lemma A.4. *Let $Z^* \in \{0,1\}^{n\times K_0}$ be a binary factor matrix with a well specified group structure such that $K_0^* \lesssim K_0^{1-\beta}$ for $\beta\in(0,1)$. Under the assumption of Theorem 4.4,*

$$E_{Z^*}\Pi\left(\left\|ZZ^T - Z^*(Z^*)^T\right\|_F^2 > \eta^2, \left\|ZZ^T + I\right\|_\infty \le C_1(K_0+1)\Big| X\right)$$

$$\le 2\exp\left(-Cp\min\left(\frac{\eta^2}{n^2K_0^2}, \frac{\eta}{nK_0}\right) + 2\log n + C_2 n K_0^{2(1-\beta)}\right),$$

for some $C, C_1, C_2 > 0$.

Proof. We prove this result using Theorem 4.1 for

$$U = \left\{\left\|ZZ^T - Z^*(Z^*)^T\right\|_F^2 > \eta^2, \left\|ZZ^T + I\right\|_\infty \le C_1(K_0+1)\right\}.$$

Using the argument in the proof of Lemma 4.1, there is a testing function ϕ, such that

$$E_{Z^*}\phi + \sup_{Z\in U} E_Z(1-\phi) \le 2\exp\left\{-Cp\min\left(\frac{\eta^2}{n^2K_0^2}, \frac{\eta}{nK_0}\right) + 2\log n\right\}.$$

Combining with the prior mass lower bound in Theorem 4.2 and Theorem 4.1, we obtain the result. □

Finally, we are ready to prove Theorem 7.1.

Proof. Without loss of generality, we assume $\left\|ZZ^T - Z^*(Z^*)^T\right\|_F \ge 1$. The case $\left\|ZZ^T - Z^*(Z^*)^T\right\|_F < 1$ implies that $\left\|ZZ^T - Z^*(Z^*)^T\right\|_F = 0$ and has been treated by Theorem 4.4. Define

$$V = \left\{\left\|ZZ^T - Z^*(Z^*)^T\right\|_F^2 > \eta^2\right\},$$

for some η to be specified later. First, we use union bound to obtain

$$\begin{aligned} E_{Z_0}\Pi(V|X) &\leq E_{Z_0}\Pi\left(V, \left\|ZZ^T+I\right\|_\infty \leq C_1(K_0+1)|X\right) \\ &\quad + E_{Z_0}\Pi\left(\left\|ZZ^T+I\right\|_\infty > C_1(K_0+1)|X\right), \end{aligned}$$

where the second term is bounded by $\exp(-C_2 p)$ according to Lemma A.3. For the first term, we bound it by

$$\begin{aligned} & E_{Z_0}\Pi\left(V, \left\|ZZ^T+I\right\|_\infty \leq C_1(K_0+1)|X\right) \\ \leq\ & \exp\left(p\|Z_0Z_0^T - Z^*(Z^*)^T\|_{\mathrm{F}}^2\right) E_{Z^*}\Pi\left(V, \left\|ZZ^T+I\right\|_\infty \leq C_1(K_0+1)|X\right) \\ & + \frac{2}{p\|Z_0Z_0^T - Z^*(Z^*)^T\|_{\mathrm{F}}^2} \\ \leq\ & 2\exp\left(-Cp\min\left(\frac{\eta^2}{n^2K_0^2}, \frac{\eta}{nK_0}\right) + 2\log n + C_2 nK_0^{2(1-\beta)} + p\|Z_0Z_0^T - Z^*(Z^*)^T\|_{\mathrm{F}}^2\right) \\ & + \frac{2}{p}, \end{aligned}$$

where the first inequality is due to Lemma A.2, and the second inequality is due to Lemma A.4 and $\|Z_0Z_0^T - Z^*(Z^*)^T\|_{\mathrm{F}}^2 \geq 1$. Choosing

$$\eta^2 = M'\frac{n^4K_0^{6-4\beta}}{p^2} + n^2K_0^2\|Z_0Z_0^T - Z^*(Z^*)^T\|_{\mathrm{F}}^4,$$

for some sufficiently large $M' > 0$, we have

$$p\min\left(\frac{\eta^2}{n^2K_0^2}, \frac{\eta}{nK_0}\right) \asymp nK_0^{2(1-\beta)} + p\|Z_0Z_0^T - Z^*(Z^*)^T\|_{\mathrm{F}}^2.$$

Then,

$$E_{Z_0}\Pi\left(V|X\right) \leq \exp\left(-C_1 nK_0^{2(1-\beta)}\right) + \exp\left(-C_2 p\right) + \frac{2}{p} \leq \frac{C_3}{p} \leq \exp\left(-C' nK_0^{2(1-\beta)}\right) + \frac{2}{p}.$$

Finally, observe that

$$\begin{aligned} V &\supset \left\{\left\|ZZ^T - Z_0(Z_0)^T\right\|_F^2 \geq M_1\left(\frac{n^4K_0^{6-4\beta}}{p^2} + n^2K_0^2\|Z_0Z_0^T - Z^*(Z^*)^T\|_{\mathrm{F}}^4\right)\right\} \\ &\supset \left\{\left\|ZZ^T - Z_0(Z_0)^T\right\|_F^2 \geq M\left(\frac{n^3K_0^{4-2\beta}}{p} + n^2K_0^2\|Z_0Z_0^T - Z^*(Z^*)^T\|_{\mathrm{F}}^4\right)\right\} \end{aligned}$$

for some $M > 0$, where the last inequality is because $\frac{n^3K_0^{4-2\beta}}{p} \gtrsim \frac{n^4K_0^{6-4\beta}}{p^2}$ under the assumption of Theorem 4.4. Hence, we obtain the desired posterior contraction for $\left\|ZZ^T - Z_0(Z_0)^T\right\|_F^2$. □

Proof of Corollary 7.1. It is sufficient to bound $||Z_0 Z_0^T - Z^*(Z^*)^T||_F^2$. By triangle inequality, we have

$$\begin{aligned} ||Z_0 Z_0^T - Z^*(Z^*)^T||_F^2 &\leq \big(||Z_0(Z_0 - Z^*)^T||_F + ||(Z_0 - Z^*)(Z^*)^T||_F\big)^2 \\ &\leq (||Z_0|| + ||Z^*||)^2 ||Z_0 - Z^*||_F^2. \end{aligned}$$

Note that Z^* is obtained by zeroing out entries in Z_0, and thus we have $||Z^*|| \leq ||Z_0||$. Since there are at most $O(n^\delta)$ entries being zeroed out, we have $||Z_0 - Z^*||_F^2 \lesssim n^\delta$. To summarize, we obtain the bound $||Z_0 Z_0^T - Z^*(Z^*)^T||_F^2 \lesssim n^\delta ||Z_0||^2$. The requirement that $(nK_0)^2 n^{2\delta} ||Z_0||^4 = o(K_0^4 n^3/p)$ leads to the condition $n^{2\delta} = o\left(\frac{nK_0^2}{p||Z_0||^4}\right)$. Thus, the proof is complete. □

B Proof of Technical Lemmas

To prove Lemma A.1, we need the following large deviation inequality.

Lemma B.1. *For $\{W_{i1}, W_{i2}\}_{i=1}^p$ from i.i.d. bi-variate normal distribution with* $\mathrm{Var}(W_{i1}) = \mathrm{Var}(W_{i2}) = 1$ *and* $\mathrm{Cov}(W_{i1}, W_{i2}) = \rho$, *we have for any $\epsilon > 0$,*

$$P\left\{\left|\frac{1}{p}\sum_{i=1}^p (W_{i1}W_{i2} - E(W_{i1}W_{i2}))\right| > \epsilon\right\} \leq \exp\big(-Cp(\epsilon \wedge \epsilon^2)\big),$$

for some $C > 0$.

Proof. Since W_{i1} and W_{i2} are from normal distribution, $W_{i1}W_{i2}$ is a sub-exponential random variable. To be specific, let us consider the case $\rho \geq 0$ without loss of generality. Then, W_{i1} and W_{i2} can be represented as

$$W_{i1} = \sqrt{\rho} Z + \sqrt{1-\rho} U, \quad W_{i2} = \sqrt{\rho} Z + \sqrt{1-\rho} V,$$

with U, V, Z i.i.d. $N(0,1)$. Then,

$$\begin{aligned} & P\{|W_{i1}W_{i2} - \rho| > t\} \\ = \; & P\left\{\left|\rho(Z^2 - 1) + \sqrt{\rho(1-\rho)}(ZU + ZV) + (1-\rho)UV\right| > t\right\} \\ \leq \; & P\left\{|\rho(Z^2-1)| > \frac{t}{3}\right\} + P\left\{|\sqrt{\rho(1-\rho)}(ZU+ZV)| > \frac{t}{3}\right\} + P\left\{|(1-\rho)UV| > \frac{t}{3}\right\} \\ \leq \; & P\left\{|Z^2 - 1| > \frac{t}{3}\right\} + P\left\{|Z(U+V)| > \frac{t}{3}\right\} + P\left\{|UV| > \frac{t}{3}\right\} \\ \leq \; & \exp(-Ct), \end{aligned}$$

for some constant $C > 0$. The last inequality above holds because $|Z^2 - 1|$, $|Z(U+V)|$ and $|UV|$ all have bounded sub-exponential norm. We have shown that $|W_{i1}W_{i2} - \rho|$ has bounded sub-exponential norm. For the case when $\rho < 0$, we can represent W_{i2} by $-\sqrt{\rho}Z - \sqrt{1-\rho}V$. By Proposition 5.16 of [36], the conclusion follows. □

Proof of Lemma A.1. Let $\frac{1}{p}XX^T = (\hat{\sigma}_{st})_{n\times n}$ and $ZZ^T + I = (\sigma_{st})_{n\times n}$. Then we have

$$P_Z\left\{\left\|\frac{1}{p}XX^T - (ZZ^T+I)\right\|_F > \epsilon\right\} = P_Z\left\{\sum_{s,t}(\hat{\sigma}_{st}-\sigma_{st})^2 > \epsilon^2\right\}$$
$$\leq \sum_{s,t} P_Z\left\{(\hat{\sigma}_{st}-\sigma_{st})^2 > \frac{\epsilon^2}{n^2}\right\} \leq \sum_{s,t} P_Z\left\{\frac{(\hat{\sigma}_{st}-\sigma_{st})^2}{\sigma_{ss}\sigma_{tt}} > \frac{\epsilon^2}{n^2||ZZ^T+I||_\infty^2}\right\}.$$

Using Lemma B.1, the above quantity can be upper bounded by

$$\sum_{s,t}\exp\left\{-Cp\min\left(\frac{\epsilon^2}{n^2||ZZ^T+I||_\infty^2}, \frac{\epsilon}{n||ZZ^T+I||_\infty}\right)\right\}$$
$$= \exp\left\{-Cp\min\left(\frac{\epsilon^2}{n^2||ZZ^T+I||_\infty^2}, \frac{\epsilon}{n||ZZ^T+I||_\infty}\right) + 2\log n\right\}.$$

This proves the first inequality. Using the same argument, we have

$$P_Z\left\{\left\|\frac{1}{p}XX^T - (ZZ^T+I)\right\|_\infty > \epsilon\right\} \leq \sum_{s,t} P_Z\left\{|\hat{\sigma}_{st}-\sigma_{st}| > \epsilon\right\}$$
$$\leq \sum_{s,t} P_Z\left\{\frac{(\hat{\sigma}_{st}-\sigma_{st})^2}{\sigma_{ss}\sigma_{tt}} > \frac{\epsilon^2}{||ZZ^T+I)||_\infty^2}\right\}$$
$$\leq \exp\left\{-Cp\min\left(\frac{\epsilon^2}{||ZZ^T+I)||_\infty^2}, \frac{\epsilon}{||ZZ^T+I)||_\infty}\right) + 2\log n\right\},$$

which proves the second inequality. □

C Date analysis using alternative methods

To compare with the real data analysis in [9] using a pIBP prior, we analyzed the same 134 breast cancer samples with the expression profiles of 300 genes and the mutation status of 11 genes with IBP prior. The resulting latent factor matrix is less sparse than that of pIBP, which offers compromised interpretability. Moreover, the reported features in [9] were not recovered by IBP prior, suggesting the integration of somatic mutations might lead to better understanding of gene expression (Supplementary Figure 6).