Phylogenetic Indian Buffet Process: Theory and Applications in Integrative Analysis of Cancer Genomics

Mengjie Chen,
Program of Computational Biology and Bioinformatics

Chao Gao,
Department of Statistics

Hongyu Zhao,
Department of Biostatistics, School of Public Health,
and Department of Genetics, School of Medicine,
Yale University, New Haven, CT 06520, USA.
Email: hongyu.zhao@yale.edu.

July 30, 2013

Abstract

By expressing prior distributions as general stochastic processes, nonparametric Bayesian methods provide a flexible way to incorporate prior knowledge and constrain the latent structure in statistical inference. The Indian buffet process (IBP) is such an approach that can be used to define a prior distribution on infinite binary features, where the exchangeability among subjects is assumed. Phylogenetic Indian buffet process (pIBP), a derivative of IBP, enables the modeling of non-exchangeability among subjects through a stochastic process on a rooted tree, which is similar to that used in phylogenetics, to describe relationships among the subjects. In this paper, we study both theoretical properties and practical usefulness of IBP and pIBP for binary factor models. For theoretical analysis, we established the posterior convergence rates for both IBP and pIBP and substantiated the theoretical results through simulation studies. As for application, we apply IBP and pIBP to data arising in the field of cancer genomics where we incorporate somatic mutations as prior information into gene expression data to study tumor heterogeneities. The results suggest that incorporating heterogeneity among subjects through pIBP may lead to better understanding of molecular mechanisms under tumor genesis and progression.

KEYWORDS: Nonparametric Bayesian, Indian Buffet Process, Latent Factor Analysis, Gene Expression, Cancer Genomics

1. INTRODUCTION

Recently nonparametric Bayesian approaches have gained popularity in machine learning and other fields to learn structural information in the data. By expressing prior distributions as general stochastic processes, nonparametric Bayesian methods provide a flexible way to incorporate prior knowledge and constrain the latent structure. The Indian buffet process (IBP) is such a stochastic process that can be used to define a prior distribution where the latent structure could be appropriately presented in the form of a binary matrix with a finite number of rows and an infinite number of columns (Griffiths and Ghahramani 2005; Knowles and Ghahramani 2011; Griffiths and Ghahramani 2011). The exchangeability among subjects is assumed in IBP, i.e, the joint probability of the subjects being modeled by the prior is invariant to permutation. In certain applications, exogenous information may suggest certain groupings of the subjects, such as studies involving normal and cancer individuals or studies involving cancer patients with different mutations. In these cases, treating all the subjects exchangeable using IBP is not appropriate. As an alternative, phylogenetic Indian buffet process (pIBP) (Miller, Griffiths and Jordan 2012) provides a flexible framework to incorporate prior structural information among the subjects for more accurate statistical inference. In pIBP, the dependency structure among subjects is captured by a stochastic process on a rooted tree similar to that used in phylogenetics. As a derivative of IBP, pIBP inherits many of the nice features of IBP including inducing sparsity, allowance of a potentially infinite number of latent factors, and data driven inference on the number of latent factors. In addition, pIBP provides an effective approach to incorporating useful information on the relationship among subjects without losing computational tractability.

Despite many successful applications of IBP and its variants in many substantive areas, as far as we know, there has not been theoretical investigation of their posterior behavior. If there is an underlying data-generating model, we expect that the posterior distributions of the model parameters from a desired inferential procedure should concentrate in the neighborhoods of the true parameter values asymptotically.

In this paper, we study both theoretical properties and practical usefulness of IBP and pIBP for binary factor models. For theoretical analysis, we established the posterior convergence rates for both IBP and pIBP and substantiated the theoretical results through simulation studies. To our

knowledge, this is the first work on the frequentist property of the posterior behaviors of IBP and pIBP. We apply IBP and pIBP to analyze cancer genomics data. Cancer research has been revolutionized by recent advances in high through-put sequencing that has enabled people to scrutinize the cancer genomes at single nucleotide resolution. Diverse types of genomics data, e.g., DNA, RNA, and epigenetic, have been generated for different tumor types (Nik-Zainal, Van Loo, Wedge, Alexandrov, Greenman, Lau, Raine, Jones, Marshall, Ramakrishna et al. 2012; Muzny, Bainbridge, Chang, Dinh, Drummond, Fowler, Kovar, Lewis, Morgan, Newsham et al. 2012; Bell, Berchuck, Birrer, Chien, Cramer, Dao, Dhir, DiSaia, Gabra, Glenn et al. 2011; TCGA 2012). These data have revealed that substantial heterogeneities exist across tumor types, across individuals within the same tumor types and even within an individual tumor. The identifications of such heterogeneities through joint analysis of different data types, e.g. somatic mutations, gene expressions, and epigenetic changes, will not only lead to better tumor subtyping, but also deepen our understanding of molecular mechanisms in tumor onset, growth, and progression. A number of computational approaches have been developed to integrate different data types for studying heterogeneity. For example, several TCGA consortium projects (Bell et al. 2011; Muzny et al. 2012; TCGA 2012), applied non-negative matrix factorization or Recursively Partitioned Mixture Model to characterize gene expression profiles, and manually integrated the somatic mutation profiles to identify tumor subtypes. Mo and colleagues have developed iCluster+ (Mo, Wang, Seshan, Olshen, Schultz, Sander, Powers, Ladanyi and Shen 2013) to jointly model genomic, epigenomic and transcriptional profiles, where somatic mutations and gene expressions were used as independent features to identify cancer subtypes. The correlation or causal relationships among different data types were not considered explicitly in these analyses.

In this paper, we propose to use a binary factor model to integrate somatic mutation and gene expression data based on a pIBP prior. Our working hypothesis is that gene expression profiles of a cancer patient may be predicted by a set of latent factors that represent distinct molecular drivers. With this hypothesis, the more similar the somatic mutation profiles are between two cancer patients, the more similar their gene expression profiles are. Therefore, we can treat available somatic mutation data from some well known driver genes as prior information using a pIBP prior specified on the latent factor matrix (elaborated in Section 6.1). By inferring the latent factors in

combination of gene expression data, we may more effectively model relationships between somatic mutations and gene expression levels, and more accurately study heterogeneities across cancer patients.

We note that pIBP echoes the phylogenetic tree representation in the cancer literature (Yates and Campbell 2012) to describe the relationships among cancer samples. More specifically, the branches in the phylogenetic tree represent the acquisitions of different somatic mutations during tumor evolution for each individual.

Our theoretical, simulation, and real data analyses show that pIBP is an attractive alternative to IBP when the subjects can be related through a tree structure based on some prior information. Moreover, even when the tree structured is mis-specified in the pIBP prior, the posterior behavior is still comparable with that of the IBP prior when the sample size is large, suggesting good robustness of the pIBP approach.

We organize the rest of the paper as follows. Section 2 presents the methodological details of the binary factor model. The definitions of IBP and pIBP are reviewed in Section 3. Section 4 presents our theoretical studies of the posterior contraction rates of IBP and pIBP. Simulation studies are carried out in Section 5. Section 6.1 introduces the somatic mutation data and describes the construction of the tree prior given somatic mutation profiles. Sections 6.2 and 6.3 present the analyses of two TCGA data sets using different prior construction strategies. Finally, Section 7 provides additional discussion about our method and results. Proofs for theoretical results are collected in Section 8.

1.1 Notations

We denote $\max(a, b)$ by $a \vee b$ and $\min(a, b)$ by $a \wedge b$. For two positive sequences $\{a_n\}$ and $\{b_n\}$, $a_n \lesssim b_n$ means there exists a $C > 0$, such that $a_n \leq C b_n$ for all n. For a matrix $A = (a_{ij})_{m\times n}$, denote its matrix Frobenius norm by $||A||_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2}$ and the sup-norm by $||A||_\infty = \max_{i,j} |a_{ij}|$. When A is a squared symmetric positive definite matrix, its spectral norm is defined as $||A|| = \lambda_{\max}(A)$. The norm $||\cdot||$, when applied on a vector, is understood to be the Euclidean norm. For a set S, denote its cardinality by $|S|$. The symbol Π stands for the prior probability distribution associated with mixture of IBP or pIBP defined in Section 3.3. The notations $\mathbb{P}$ and $\mathbb{E}$ are used as generic

probability or expectation when the distribution is clear from the context.

2. BINARY FACTOR MODEL

For gene expression data, let $X = (x_{ij})_{n\times p}$ denote the observed data matrix, where each of the n rows represents one individual and each of p columns represents one gene. We hypothesize that gene expression profiles can be captured by distinct molecular drivers. We model the effects of these drivers by a set of latent factors Z through the following model:

$$X = ZA + E,$$

where $Z = (z_{ik})_{n\times K}$ is a binary factor matrix, and $A = (a_{kj})_{K\times p}$ is the loading matrix. The status of z_{ik}, which takes value of 1 or 0, indicates the presence or the absence of the k-th factor in the i-th individual. The value of a_{kj} weighs the contribution of the j-th gene to the k-th factor. We assume each entry of $E = (e_{ij})_{n\times p}$ follows $N(0, \sigma_X^2)$ independently. Let each entry of A follows i.i.d. $N(0, \sigma_A^2)$, and A is independent of E. Conditioning on A, we assume $(X|A)$ follows a matrix normal distribution with mean ZA. Integrating out A with respect to its distribution, each column of X follows

$$(x_{1j}, x_{2j}, ..., x_{nj})^T \sim N(0, \sigma_A^2 ZZ^T + \sigma_X^2 I), \tag{1}$$

independently for $j = 1, ..., p$. Compared with formula (49) in Griffiths and Ghahramani (2005), (1) is easier to work with to model X, and it also shows the covariance structure across individuals imposed by the binary factor model.

2.1 Feature Similarity Matrix ZZ^T

We name matrix ZZ^T the feature similarity matrix because of its important statistical meaning as reflected in (1). We denote it by $ZZ^T = (\gamma_{ij})_{n\times n}$. Each row/column of this matrix ZZ^T describes the feature similarity between a particular individual and the other $n-1$ individuals. Notice that

$$\gamma_{ij} = \sum_{k=1}^{K} z_{ik} z_{jk} = |\{k : z_{ik} = z_{jk} = 1\}|.$$

Thus, the diagonal element γ_{ii} denotes the number of factors possessed by the i-th individual, whereas the off-diagonal entry γ_{ij} is the number of the factors shared between the i-th and j-th individuals. In short, the feature similarity matrix ZZ^T characterizes the latent driver sharing

structure among samples. For the i-th individual, we define $d_i = \sum_{j \neq i} \gamma_{ij}$ as its degree. When we have a group structure among the samples, the individual with the highest degree has the most shared factors among a group. That particular individual is a representative prototype for that group.

3. TREE STRUCTURED INDIAN BUFFET PROCESS PRIOR

To pursue a full Bayesian approach, we put prior on the triple $(Z, \sigma_A^2, \sigma_X^2)$. Let π be the prior on (σ_A^2, σ_X^2). The choice of π is not essential because when n and p are large, the prior effect on the parametric part (σ_A^2, σ_X^2) is negligible. In contrast, the prior on the binary matrix Z is important. Since we do not specify the number of columns K in advance, the potential number of parameters in Z is infinite. It is well-known that when the number of parameters grows with the sample size, Bayesian method is no longer guaranteed to be consistent (Diaconis and Freedman 1986). Thus, the choice of prior on Z is important. Another issue is the identifiability of the model. As we observe in the model representation (1), the order of the columns of Z is not identifiable. In other words, we cannot tell the first factor from the second. Instead of specifying the prior on Z, we specify the prior on the equivalent class $[Z]$, where $[Z]$ denotes the collection of matrices Z which are equivalent by reordering the columns.

We describe two priors on $[Z]$ in this section, the Indian buffet process proposed by Griffiths and Ghahramani (2005), and its tree-structured generalization, the phylogenetic Indian buffet process proposed by Miller et al. (2012). Both are priors on sparse infinite binary matrices.

3.1 Indian Buffet Process

We describe the Indian buffet process (IBP) on $[Z]$ by its stick-breaking representation derived in Teh, Görür and Ghahramani (2007). First draw an infinite sequence $\{p_k\}$ by

$$v_k \sim \text{Beta}(\alpha, 1), \quad \text{i.i.d.}, \quad \text{and} \quad p_k = \prod_{i=1}^{k} v_i. \tag{2}$$

Given $\{p_k\}$, z_{ik} is drawn independently from a Bernoulli distribution with parameter p_k for $i = 1, ..., n$ and $k = 1, 2, 3, ...$ The final matrix Z drawn in this way has dimension $n \times K^+$, where K^+ is the number of nonzero columns and is finite with probability 1. The IBP prior on $[Z]$ is the image measure induced by the equivalence map $Z \mapsto [Z]$.

3.2 Phylogenetic Indian Buffet Process

The phylogenetic Indian buffet process (pIBP) also starts with drawing $\{p_k\}$ as in (2). Different from IBP, given p_k, the entries of the k-th column of Z are not i.i.d. in pIBP. Their dependency structure is captured by a stochastic process on a rooted tree similar to the models used in phylogenetics (Miller et al. 2012). The n individuals are modeled as leaves of the tree. The total edge length from the root to any leaf is 1. Conditioning on p_k, we describe the generating process of the k-th column of Z. First, assign 0 to the root of the tree. Along any path from the root to a leaf, let the value change to 1 along any edge of length t with probability $1 - \exp(-\gamma_k t)$, where $\gamma_k = -\log(1 - p_k)$. Once the value has changed to 1 along any path from the root, all leaves below that point are assigned value 1. The pIBP prior is defined to be the image measure on $[Z]$.

3.3 Mixture of IBP and pIBP

Both IBP and pIBP are determined by the hyper-parameter α, which can be tuned in practice. In this paper, we pursue a full-Bayesian approach, and thus put a $\Gamma(1, 1)$ prior on α for both IBP and pIBP. Thus, the final prior on the equivalent class $[Z]$ is a mixture of IBP or pIBP.

4. POSTERIOR CONTRACTION RATES OF IBP AND PIBP

In this section, we establish the posterior convergence of both mixture of IBP and mixture of pIBP and characterize their difference by different convergence rates. Such theoretical comparisons are interesting because IBP can be viewed as a special case of pIBP with a default tree. These results will illustrate the impacts of tree structure imposed by the prior.

4.1 Convergence of the Feature Similarity Matrix

We define the triple $(Z_0, \sigma^2_{A,0}, \sigma^2_{X,0})$ to be the true parameter generating the data matrix X, where Z_0 is an $n \times K_0$ binary matrix and K_0 is the number of factors. For the sake of clearer presentation, we assume $\sigma^2_{A,0} = \sigma^2_{X,0} = 1$, so that the only unknown parameter is Z_0. The generalization to unknown variances is covered in Section 8. Let Π be the mixture of IBP or pIBP prior on $[Z]$. Notice that the matrix ZZ^T does not depend on the order of columns of Z, so that we have $ZZ^T = [Z][Z]^T$. We consider the posterior convergence in the sense of

$$\Pi\Big(||ZZ^T - Z_0 Z_0^T||_F^2 \leq M\epsilon_{n,p}^2 \Big| X\Big) \longrightarrow 1, \quad \text{in } P_{Z_0}\text{-probability}, \tag{3}$$

with some sequence $\epsilon_{n,p}$ and constant $M > 0$. We choose to study the posterior convergence in terms of the feature similarity matrix ZZ^T because of both the identifiability issue and statistical interpretation described in Section 2.1.

4.2 A General Method

The theory of Bayesian posterior consistency was first studied by Schwartz (1965). She proposed the Kullback-Leibler property of the prior and the testing argument to prove weak consistency in the parametric case. The first nonparametric posterior consistency result was obtained by Barron (1988), where the idea of testing on the essential support of the prior is used. Later, the same argument was modified to achieve rate of convergence by Ghosal, Ghosh and van der Vaart (2000). In the current setting of binary factor model, we propose the following general method to prove posterior rate of convergence. For each Z, denote P_Z to be the generating process of the model $X = ZA + E$. We have the following theorem.

Theorem 4.1. *For any measurable set U, and any testing function ϕ, we have*

$$P_{Z_0}\Pi\Big(U|X\Big) \leq P_{Z_0}\phi + \frac{1}{\Pi\Big(||ZZ^T - Z_0Z_0^T||_F^2 = 0\Big)} \sup_{Z\in U} P_Z(1-\phi). \tag{4}$$

The theorem can be viewed as a discrete version of the Schwartz theorem (Schwartz 1965). We take advantage of the discrete nature of the problem, thus avoiding calculating the prior mass of the Kullback-Leibler neighborhood of P_{Z_0}. We specify U to be

$$U = \left\{||ZZ^T - Z_0Z_0^T||_F^2 > M\epsilon_{n,p}^2\right\}.$$

Thus, in order to obtain (3), it is sufficient to upper bound the right hand side of (4). This can be done by lower bounding $\Pi\Big(||ZZ^T - Z_0Z_0^T||_F^2 = 0\Big)$ and constructing a test for $H_0 : Z = Z_0$ and $H_1 : Z \in U$ with appropriate type 1 and type 2 error bounds. Such test is guaranteed by the following lemma.

Lemma 4.1. *For any $\epsilon_{n,p} > 0$, there is a testing function ϕ, such that the testing error $P_{Z_0}\phi + \sup_{\{||ZZ^T - Z_0Z_0^T||_F^2 > M\epsilon_{n,p}^2\}} P_Z(1-\phi)$ is upper bounded by*

$$\exp\left\{-C\sqrt{M}p\min\left(\frac{\epsilon_{n,p}^2}{n^2K_0^2}, \frac{\epsilon_{n,p}}{nK_0}\right) + 2\log n\right\} + \exp\Big(-Cp^{1/2} + 2\log n\Big),$$

for some universal constant $C > 0$.

Therefore, it is sufficient to lower bound $\Pi\Big(||ZZ^T - Z_0Z_0^T||_F^2 = 0\Big)$ to obtain (3). The posterior convergence rate is fully determined by the prior concentration.

4.3 Two-Group Tree and Factor Decomposition

In this section, we study a special but representative pIBP structure, the two-group tree. Let n individuals be labeled by $\{1, 2, ..., n\}$. Without loss of generality, we assume n is even. Let $\{1, ..., n\} = S_1 \cup S_2$, where $S_1 = \{1, 2, ..., [n/2]\}$ and $S_2 = \{[n/2]+1, ..., n\}$. The tree induced by the two-group structure (S_1, S_2) has one root, two group nodes and n leaves. The two group nodes are connected with the root by two edges of length $\eta \in (0, 1)$. Then, the i-th group node is connected with each member of S_i by an edge of length $1 - \eta$, where $i = 1, 2$. The parameter η is the strength of the group structure imposed by the prior Π. When, $\eta = 0$, pIBP reduces to IBP.

[Figure 1 about here.]

Our theory covers three cases. The first case is the IBP prior, with no group structure specified in the prior. The second case is the two-group pIBP prior with group structure correctly specified. The third case is the the two-group pIBP prior with group structure mis-specified. Let Z_0 have K_0 columns, representing K_0 factors. Given the two-group structure (S_1, S_2) by the prior Π, we have the following decomposition

$$K_0 = K_{01} + K_{02} + K_0^*, \tag{5}$$

where K_{01} is the number of factors unique to S_1, K_{02} is the number of factors unique to S_2, and K_0^* is the number of factors shared across S_1 and S_2. The decomposition (5) is determined by both the structure of Z_0 and the prior Π. It characterizes how well the group structure is specified compared with the truth Z_0 (see Figure 1). Generally speaking, the smaller the K_0^* is, the better the group structure is specified by Π.

4.4 Prior Concentration

Theorem 4.2. *The probability* $\Pi\Big(||ZZ^T - Z_0Z_0^T||_F^2 = 0\Big)$ *can be lower bounded by*

$$A^{-(K_0+\kappa)(K_0-K_0^*+1)} \exp\left(- Cn(K_0^* + \kappa)^2 - Cn(1-\eta) - Cn(1-\eta)\frac{K_0 - K_0^*}{(4/3)^{(K_0^*+\kappa)}}\right),$$

for any $\kappa > 0$. *The constants* $A > 1$ *and* $C > 0$ *are universal.*

Theorem 4.2 provides an explicit characterization of the prior concentration rate as a function of K_0, K_0^* and η. The prior concentration rate directly determines the posterior convergence rate according to Theorem 4.1 and Lemma 4.1. In the following, we consider $\eta = 0$ and $\eta \in (0, 1)$, separately.

1. $\eta = 0$. In this case, pIBP and IBP are equivalent. The prior does not impose any group structure. Thus, in the decomposition (5), we have $K_0^* = K_0$. By letting $\kappa = 0$, Theorem 4.2 can be written as

$$\Pi\Big(||ZZ^T - Z_0 Z_0^T||_F^2 = 0\Big) \geq A^{-K_0} \exp\Big(- Cn(K_0^2 + 1)\Big). \tag{6}$$

   The prior concentration rate for IBP in (6) is the benchmark for us to compare IBP with pIBP in various situations.

2. $\eta \in (0, 1)$. In this case, the tree structure plays a role in the prior. In practice, $\eta = 1/2$ is often used to characterize moderate group structure belief in the prior. We say the group structure is effectively specified if $K_0^* \lesssim K_0^{1-\beta}$ for some $\beta \in (0, 1)$. In this case, the result of Theorem 4.2 can be optimized for $k = K_0^* + \kappa$ for any $\kappa > 0$. That is, for n sufficiently large, we have

$$\Pi\Big(||ZZ^T - Z_0 Z_0^T||_F^2 = 0\Big) \geq \exp\left(- 2Cn \max_{k \geq K_0^*} \Big(k^2 \vee \frac{(1-\eta)K_0}{(4/3)^k}\Big)\right), \tag{7}$$

   which is lower bounded by

$$\exp\Big(- 2CnK_0^{2(1-\beta)}\Big).$$

   This rate is superior to (6). Thus, pIBP is advantageous over IBP as long as the tree structure captures any group-specific features in the sense that $K_0^* = o(K_0)$. On the other hand, the group structure is mis-specified if $K_0^* = K_0$. In this case, we reduce to (6), so that

$$\Pi\Big(||ZZ^T - Z_0 Z_0^T||_F^2 = 0\Big) \geq A^{-K_0} \exp\Big(- Cn(K_0^2 + 1)\Big).$$

   Thus, a mis-specified tree structure does not compromise the results. As to the possibility that this is due to a loose bound in (6), by scrutinizing the proof, we found that the slack is at most in a constant level independent of (n, K_0, K_0^*). Thus, the prior concentrations of a pIBP with a mis-specified tree and of the IBP are essentially the same.

4.5 Posterior Convergence Rates

Combining Theorem 4.1, Lemma 4.1 and Theorem 4.2, we can derive the posterior convergence rates in the sense of (3) for both IBP and pIBP.

Theorem 4.3. *For the mixture of IBP prior* Π *on* $[Z]$*, let* Z_0 *be the true factor matrix. Then, for the binary factor model, there exists* $M > 0$ *and* $C_1 > 0$*, such that*

$$P_{Z_0}\Pi\Big(||ZZ^T - Z_0Z_0^T||_F^2 \leq M\big(\frac{K_0^4 n^3}{p} \vee \frac{K_0^6 n^4}{p^2}\big)\Big|X\Big) \geq 1 - \exp\Big(-C_1\big(p^{1/2} - \log n\big)\Big).$$

Theorem 4.4. *For the mixture of pIBP prior* Π *on* $[Z]$ *with* $\eta \in (0,1)$*, let* Z_0 *be the true factor matrix. When* $K_0^* \lesssim K_0^{1-\beta}$ *for* $\beta \in (0,1)$*, for the binary factor model, there exists* $M > 0$ *and* $C_1 > 0$*, such that*

$$P_{Z_0}\Pi\Big(||ZZ^T - Z_0Z_0^T||_F^2 \leq M\big(\frac{K_0^{4-2\beta} n^3}{p} \vee \frac{K_0^{6-4\beta} n^4}{p^2}\big)\Big|X\Big) \geq 1 - \exp\Big(-C_1\big(p^{1/2} - \log n\big)\Big).$$

When $K_0^* = K_0$*, there is* $M > 0$*, such that*

$$P_{Z_0}\Pi\Big(||ZZ^T - Z_0Z_0^T||_F^2 \leq M\big(\frac{K_0^4 n^3}{p} \vee \frac{K_0^6 n^4}{p^2}\big)\Big|X\Big) \geq 1 - \exp\Big(-C_1\big(p^{1/2} - \log n\big)\Big).$$

The above two theorems establish the rates of convergence for the posterior distributions of IBP and pIBP, respectively. As long as $\big(\log n\big)^2 = o(p)$, the posterior probabilities converge to 1 in expectation under the true model. Compared with the rate of IBP in Theorem 4.3, when the tree structure is effectively specified, the rate of pIBP in Theorem 4.4 is faster with a factor of $K_0^{2\beta}$ or $K_0^{4\beta}$, which depends on the asymptotic regimes of (n, p, K_0). Such difference is significant if the number of features K_0 is large. Moreover, Theorem 4.4 also suggests that even when the tree structure of pIBP is mis-specified, the rate of convergence is the same as that of IBP, implying the robust property of pIBP. Although our theoretical study is carried out in the simple two-group structure model, similar conclusions may be obtained under a more complicated structural assumption using the same method.

5. SIMULATION STUDIES

In this section, we perform simulations to evaluate the performance of IBP and pIBP. We implemented the MCMC algorithm proposed in (Miller et al. 2012) to perform posterior inference of latent factor matrix Z. In the algorithm, the sampling process on the tree structure is expressed as a graphical model, where the prior probabilities can be calculated efficiently by a sum-product algorithm. All the parameters σ_A, σ_X, α and $p = \{p_k\}$ (marginal probabilities of a latent feature equaling 1) are sampled as part of the overall MCMC procedure.

In the first simulation, we evaluated the performance of IBP, pIBP with an appropriate tree structure, and pIBP with a mis-specified tree structure (mispIBP). We constructed a set of samples with a clear subgroup structure on Z_0. Specifically we simulated data with eight subgroups characterized by six latent factors as an illustration of Z_0 shown in Figure 2. Twelve models presented in Table 1 are considered. For each model, we generated an $n \times p$ matrix $X = Z_0 A + E$ with $(\sigma_{A,0}, \sigma_{X,0}) = (1, 0.5)$. For IBP, we let $\eta = 0$ so that pIBP is equivalent to IBP. For pIBP, we let $\eta = 0.8$ and the proper tree structure is given. For mispIBP, we let $\eta = 0.5$ and the prior is a mis-specified tree with samples within a subtree assigned to different groups. Estimation error on Z is evaluated in terms of the Frobenius norm of the feature similarity matrix $\sqrt{n^{-1}||ZZ^T - Z_0 Z_0^T||_F^2}$. We further evaluated the latent structure recovery by the number of estimated latent features.

Generally, reported twelve models represent two scenarios: the small p scenario and the large p scenario. Remember in our setting, the larger the value of p is, the more accurately we can recover the latent features. In the models with a small p ($p = 30$ and 20), the information from data is limited and the inference relies more heavily on the prior information. We found pIBP performs better than the other two methods in both cases. Although in the small p scenario all the methods overestimate the number of latent factors, pIBP prior effectively controls the introduction of new factors. Besides, mispIBP has comparable performance with IBP, implying that pIBP is robust to mis-specified tree structure. The simulation results substantiate the conclusions we have from Theorem 4.3 and Theorem 4.4. In the models with large p ($p = 100$ and 200), there is adequate information from the data and the priors play a less role. Inferences using different priors lead to similar results.

[Figure 2 about here.]

[Table 1 about here.]

In the second simulation, we used the similarity data to construct the pIBP prior. Nine models presented in Table 2 are considered. For each model, we generated an $n \times K_0$ binary matrix Z_0 with 4 columns sampled from a Bernoulli (0.3) and 5 columns with fixed structure. For IBP, no prior of the group structure is given. For pIBP, the tree prior is adapted from the dendrogram based on a hierarchical clustering analysis on Z_0 (see Figure 3). For mispIBP, the tree prior was constructed in the same way as pIBP but using a random permutation of Z_0 on rows in the clustering. In this setting, mispIBP represents totally incorrect information. We simulated 40 independent datasets for each model. Similar as the previous simulation, we evaluated the performance by $\sqrt{n^{-1}||ZZ^T - Z_0Z_0^T||_F^2}$ and the number of estimated latent features (Table 2). In our simulation, when p is small, pIBP outperforms IBP in all cases. In our analysis, we constructed our prior from the true knowledge on Z_0. In practice, such trees need to be constructed from external sources. The results on the mis-specified tree represent a worst case scenario. When p is adequately large ($p = 60$ in this setting), the inference is less influenced by the prior information.

[Figure 3 about here.]

[Table 2 about here.]

6. APPLICATIONS OF PIBP IN THE INTEGRATIVE CANCER GENOMICS ANALYSIS

6.1 Introduction of Somatic Mutation Data and Strategies to Construct the Tree Prior

We consider studies on a specific cancer type/subtype, which collects somatic mutations from whole exome sequencing and gene expressions either from sequencing or microarrays for each sample. Somatic mutations can either be more narrowly defined as single nucleotide changes and small insertions/deletions, or more broadly defined to include changes at the copy number level. We denote the detected somatic mutations for a group of samples by a binary matrix $S = (s_{il})_{n\times m}$, with s_{il} indicating the mutation status of the l-th gene on the i-th individual, as an external resource to construct the tree prior. When subclonality information is available, s_{il} may be expressed as a continuous measure between 0 and 1, representing the percentage of the cells harboring mutations at the l-th gene.

As for using a tree structure to express the relationships of individuals using the somatic mutation data, we consider the following three approaches:

1. **Logic Tree.** The logic tree prior is constructed as a logic tree based on the presence/absence of a set of somatic mutations. In this case, each node represents the status of a specific mutation, e.g. NF1+ or NF1- to denote the presence or absence of mutations on gene NF1. The order of mutation acquisitions could be incorporated into the tree structure. An example is shown in Figure 4.

2. **Dendrogram Tree.** The dendrogram tree prior is adapted from the dendrogram tree of a hierarchical clustering on the somatic profiles $S = (s_{il})_{n\times m}$. In such a tree, the non-leaf nodes have no explicit meaning but represent a local cluster of individuals. When the order of mutation acquisitions and the effects of specific mutations are unknown, the dendrogram tree provides a measure of the overall similarities between individuals. In practice, we perform hierarchical clustering analysis only using frequently observed mutations (e.g. observed in at least 5% of the samples). An example is shown in Figure 5.

3. **Event Tree.** The event tree prior is customized with each node representing any remarkable somatic event or any combination of somatic events. Subclonality information could be incorporated through this setting.

6.2 Analysis of TCGA GBM Data

We first consider the TCGA GBM Level 3 dataset reported in (McLendon, Friedman, Bigner, Van Meir, Brat, Mastrogianakis, Olson, Mikkelsen, Lehman, Aldape et al. 2008) (downloaded from cBio (Cerami, Gao, Dogrusoz, Gross, Sumer, Aksoy, Jacobsen, Byrne, Heuer, Larsson et al. 2012)). This dataset contains complete somatic mutations, copy number variations, and normalized gene expression measurements for 91 glioblastoma (GBM) samples. In the pathogenesis of diffuse gliomas, primary GBM typically arises through several ways that activate pro-growth pathways and suppress tumor suppressors. Primary GBM could be further characterized into subtypes based on their transcriptional signatures. These subtypes are closely associated with the mutation status of TP53, PTEN and NF1 (Brennan 2011). In contrast, the initial acquisition of IDH1 and the associated G-CIMP DNA methylation pattern appear to be a common early event for the development of

secondary GBM. Secondary GBM are more malignant and prone to metastasis. Thus identification of gene markers associated with IDH1 mutation help with grade classification and is of great importance in both diagnosis and prognosis for GBM. Incorporating such knowledge, we constructed a logic tree prior with the mutation status of IDH1, TP53, PTEN, and NF1 (Figure 4), where "+" represents somatic level abnormality with either mutation or significant copy number aberration, and "-" indicates no observed somatic aberration. We removed samples not harboring any of these mutations, leading to a subset of 61 samples. For expression data, genes with the top 1000 MAD (Median Absolute Deviation) across 61 samples were kept and centered. We ran 10 parallel MCMC chains. No substantial difference was observed across runs and we chose the one with the largest posterior probability as the final result. No substantial difference was observed from different runs. Figure 4 shows the inferred latent feature matrix Z and feature similarity matrix ZZ^T sampled from the posterior of pIBP. Given inferred Z, we estimated the loading matrix A by ridge regression (Griffiths and Ghahramani 2005):

$$\mathbb{E}\Big(A|X,Z\Big) = \Big(Z^TZ + \frac{\sigma_X^2}{\sigma_A^2}I\Big)^{-1}Z^TX.$$

[Figure 4 about here.]

We found samples with IDH1 mutation formed a cluster on ZZ^T and shared more features compared to samples without IDH1 mutation. IDH1 positive samples are characterized by their depletion of 1st, 12th, 14th and 17th latent features. We further investigated the latent features by examining the genes with top loadings. Examination on the 1st latent feature revealed its association with the expression levels of genes including EGFR, HERPUD2, NT5C2, TMEFF2, TMEM100, GALNT13, BCL2L12, METTL7B, DDRGK1, DOG1(TMEM161A), PDCD5, and SERPINE1. Among these, aberrational expression on EGFR is an known frequent event associated with classical GBM (Verhaak, Hoadley, Purdom, Wang, Qi, Wilkerson, Miller, Ding, Golub, Mesirov et al. 2010). HERPUD2 is coamplified with EGFR and DDRGK1 is coupled with NF-κB Signaling, which may contribute to the tumorigenesis process (Nord, Hartmann, Andersson, Menzel, Pfeifer, Piotrowski, Bogdan, Kloc, Sandgren, Olofsson et al. 2009; Wu, Lei, Mei, Tang, Li, Wu, Lei, Mei, Tang and Li 2010). METTL7B is up-regulated by TP53 mutants (Neilsen, Noll, Suetani, Schulz, Al-Ejeh, Evdokiou, Lane, Callen et al. 2011). TMEM100 and TMEFF2 exhibit antipro-

liferative effects or potential tumor suppressor activity, the turnover of which may indicate the oncogenesis (Gery, Sawyers, Agus, Said, Koeffler et al. 2002; Frullanti, Colombo, Falvella, Galvan, Noci, De Cecco, Incarbone, Alloisio, Santambrogio, Nosotti et al. 2012). PDCD5 plays an important apoptosis accelerating role in cells undergoing apoptosis, the decreased expression of which has been detected in various human carcinomas (Chen, Wang, Ma and Chen 2006). NT5C3, GALNT9, BCL2L12, DOG1, and SERPINE1 have been proposed as prognostic markers in other tumor types (acute myeloid leukemia, neuroblastoma, breast cancer, gastrointestinal stromal tumors) (Berois, Gattolliat, Barrios, Capandeguy, Douc-Rasy, Valteau-Couanet, Bénard and Osinaga 2013; Jordheim, Nguyen-Dumont, Thomas, Dumontet and Tavtigian 2008; West, Corless, Chen, Rubin, Subramanian, Montgomery, Zhu, Ball, Nielsen, Patel et al. 2004; Samarakoon, Higgins, Higgins and Higgins 2009; Thomadaki, Talieri, Scorilas et al. 2007). The 12th feature is characterized by its association with EGFR and MET. The 14th feature is characterized by CDKN2A, a gene on the RB pathway frequently inactivated in GBM, and RFAP1, an androgen receptor target gene in prostate cancer (Jariwala, Prescott, Jia, Barski, Pregizer, Cogan, Arasheben, Tilley, Scher, Gerald et al. 2007). The 17th feature is characterized by ALKBH3, which contributes to cell survival in several cancer types (Tasaki, Shimada, Kimura, Tsujikawa and Konishi 2011). Since the acquisition of IDH1 is a remarkable event for the evolution of a low grade glioma into secondary GBM, these genes are potential prognostic markers for grade classification in GBM.

6.3 Analysis of TCGA BRCA Data

We also analyzed the TCGA BRCA Level 3 dataset generated by (TCGA 2012) (downloaded from cBio (Cerami et al. 2012)) using a different tree construction strategy. We focused on 134 samples categorized as HER2 or Basal-like subtypes. Among these two subtypes, HER2 subtype is relatively well characterized and has effective clinical treatments. The basal-like subtype, which is also known as triple-negative breast cancers (TNBCs, lacking expression of ER, progesterone receptor (PR) and HER2), is poorly understood, with only chemotherapy as the main therapeutic option (TCGA 2012). Characterization of the basal-like subtype at the molecular level has important clinical implications. We built a tree prior from the dendrogram of a hierarchical clustering analysis with the frequent mutations in breast cancer including AKT1, CDH1, GATA3, MAP3K1, MLL3, PIK3CA, PIK3R1, PTEN, RUNX1 and TP53. For expression data, genes having top 300 MAD

across samples were kept and centered. We ran 10 MCMC chains. No substantial difference was observed across runs and we chose the one with largest posterior probability as the final result. Figure 5 shows the input tree prior, subtype information and the inferred latent feature matrix Z.

In our samples, the basal-like and HER2 samples display different and almost complementary patterns in their possession of the first two features. 74 of 81 Basal-like samples exhibit the first feature and 79 of 81 are deplete with the second feature. In contrast, 43 of 53 HER2 samples are deplete with the first feature and 31 of 53 exhibit the second feature. For the first feature, the top 10 genes with the largest loadings include MRPL9, PUF60, SCNM1, EIF2C2, BOP1, MTBP, DEDD, PHF20L1, HSF1 and HEATR1. Among these, BOP1 is involved in ribosome biogenesis and contributes to genomic stability, deregulation of which leads to altered chromosome segregation (Killian, Sarafan-Vasseur, Sesboüé, Le Pessot, Blanchard, Lamy, Laurent, Flaman and Frébourg 2006); MTBP inhibits cancer metastasis by interacting with MDM2 (Chène 2003); DEDD interacts with PI3KC3 to activate autophagy and attenuate epithelialmesenchymal transition in cancer (Lv, Wang, Xue, Hua, Mu, Lin, Yan, Lv, Chen and Hu 2012); and HSF1 has been proposed as a predictor of survival in breast cancer (Van De Vijver, He, van't Veer, Dai, Hart, Voskuil, Schreiber, Peterse, Roberts, Marton et al. 2002). EIF2C2, PUF60 and PHF20L1 have been reported as prognostic markers in ovarian cancer (Ramakrishna, Williams, Boyle, Bearfoot, Sridhar, Speed, Gorringe and Campbell 2010; Wrzeszczynski, Varadan, Byrnes, Lum, Kamalakaran, Levine, Dimitrova, Zhang and Lucito 2011), which is consistent with the recent discovery that basal-like breast tumours with high-grade serous ovarian tumours share many molecular commonalities (TCGA 2012). These basal-like specific genes may potentially become novel therapeutic targets or prognostic markers. For the second feature, the top 10 genes with the largest loadings include STARD3, MED1, PSMD3, GRB7, ORMDL3, WIPF2, CASC3, RPL19, SNF8 and AMZ2. Among these, overexpressions of STARD3, PSMD3, GRB7, CASC3 and RPL19 have been reported in HER2-amplified breast cancer cell lines (Arriola, Marchio, Tan, Drury, Lambros, Natrajan, Rodriguez-Pinilla, Mackay, Tamber, Fenwick et al. 2008); MED1 is required for estrogen receptor-mediated gene transcription and breast cancer cell growth (Zhang, Jiang, Xu, Zhang, Zhang, Jiang, Xu and Zhang 2011). As revealed by principal component analysis based on gene expression (Figure 5), these genes weighing high on first two latent features have discriminating power on Basal-like and HER2 samples.

[Figure 5 about here.]

Furthermore, we found that the status of the fifth and sixth features was strongly associated with disease recurrence in our samples as revealed by survival analysis (Figure 6 shows the Kaplan -Meier plot). Samples with the fifth feature have a higher probability of recurrence than those without it, with a p-value of 0.0068, whereas samples without the sixth feature have a higher probability of recurrence than those with it, with a p-value of 0.00084. Examinations of the loadings on these two features identified RMDN1, ARMC1, TMEM70, VCPIP1, TCEB1, MTDH, EBAG9, MRPL13, UBE2V2, FAM91A1 and RRS1 on the fifth feature and TRIM11, COMMD5, PYCRL, TIGD5, MRPL55, LSM1, SETDB1, CNOT7, PROSC, DEDD and HSF1 on the sixth feature. Among these, the prognosis significance of some has been discussed before, for example, MTDH activation by 8q22 genomic gain promotes chemoresistance and cetastasis of poor-prognosis breast cancer (Hu, Chong, Yang, Wei, Blanco, Li, Reiss, Au, Haffty and Kang 2009); EBAG9 (RCAS1) is associated with ductal breast cancer progression (Rousseau, Têtu, Caron, Malenfant, Cattaruzzi, Audette, Doillon, Tremblay and Guérette 2002). The others genes may serve as candidate tumor progression markers.

In comparison, 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 (See Supplementary Figure 1). Moreover, the above reported features were not recovered by IBP prior, suggesting the integration of somatic mutations might lead to better understanding of gene expression. We also applied iCluster on the same dataset. iCluster identified three latent classes without pinpointing any interesting genes (See Supplementary Figure 2). The latent factor model with pIBP prior benefits from a richer model representation and recovers the latent structure of the same datasets with more details.

[Figure 6 about here.]

7. DISCUSSION

In this paper, we established the posterior convergence rates for both IBP and pIBP under the binary factor model setting. Our theoretical framework can be applied to evaluate posterior be-

haviors of IBP and pIBP under various settings. Compared with IBP prior, both our theory and simulation results suggest pIBP is advantageous when it incorporates information on the dependency structure among subjects. Even when the tree structure is mis-specified, pIBP performs comparably with IBP in their posterior rates of convergence. Computationally, algorithmic complexity of pIBP is only slightly worse than that of IBP (Miller et al. 2012). Thus, pIBP is preferred whenever information on the latent structure is available.

As an application example, we highlighted the application of pIBP in cancer genomics to incorporate somatic mutation information into gene expression analysis. We explored two tree construction strategies with different biological assumptions. One assumes grouping structure is driven by a particular set of driver mutations (in our example, IDH1). The other assumes the overall similarity on somatic mutation profiles leads to the similarity on the gene expression profiles. The precise mutation acquisition information is not considered in either case. However, with the maturation of sequencing technology, the tumor evolution process will become more evident at clonal and subclonal level in the near future. We anticipate this information will be integrated into downstream analysis through pIBP and other strategies.

8. PROOFS

For the sake of clarity, we write ϵ for $\epsilon_{n,p}$, with the dependency on n and p being implicit. We will first prove some preparatory lemmas. Then, we will prove Theorem 4.1 and Lemma 4.1. The heart of the proof lies in Theorem 4.2. We will also generalize the result to the case where (σ_A^2, σ_X^2) are unknown.

8.1 Preparatory lemmas

Lemma 8.1. *For any $\epsilon > 0$,*

$$P_{Z_0}\left\{\left\|\frac{1}{p}XX^T - (Z_0Z_0^T + I)\right\|_F > \epsilon\right\} \leq \exp\left\{-Cp\min\left(\frac{\epsilon^2}{n^2K_0^2}, \frac{\epsilon}{nK_0}\right) + 2\log n\right\}$$

Proof. Let $\frac{1}{p}XX^T = (\hat{\sigma}_{st})_{n\times n}$ and $Z_0Z_0^T + I = (\sigma_{st})_{n\times n}$. Then we have

$$\begin{aligned}
& P_{Z_0}\left\{\left\|\frac{1}{p}XX^T - (Z_0Z_0^T + I)\right\|_F > \epsilon\right\} \\
= \; & P_{Z_0}\left\{\sum_{s,t}(\hat{\sigma}_{st} - \sigma_{st})^2 > \epsilon^2\right\} \\
\leq \; & \sum_{s,t} P_{Z_0}\left\{(\hat{\sigma}_{st} - \sigma_{st})^2 > \frac{\epsilon^2}{n^2}\right\} \\
\leq \; & \sum_{s,t} \exp\left\{-Cp\min\left(\frac{\epsilon^2}{n^2K_0^2}, \frac{\epsilon}{nK_0}\right)\right\} \\
= \; & \exp\left\{-Cp\min\left(\frac{\epsilon^2}{n^2K_0^2}, \frac{\epsilon}{nK_0}\right) + 2\log n\right\},
\end{aligned}$$

where we have used Corollary 5.35 of Vershynin (2010). □

Lemma 8.2. *For any $\epsilon > 0$,*

$$P_{Z_0}\left\{\left\|\frac{1}{p}XX^T - (Z_0Z_0^T + I)\right\|_\infty > \epsilon\right\} \leq \exp\left\{-Cp\min\left(\frac{\epsilon^2}{K_0^2}, \frac{\epsilon}{K_0}\right) + 2\log n\right\}.$$

Proof. Using the same notation as the previous lemma, we have

$$\begin{aligned}
& P_{Z_0}\left\{\left\|\frac{1}{p}XX^T - (Z_0Z_0^T + I)\right\|_\infty > \epsilon\right\} \\
\leq \; & \sum_{s,t} P_{Z_0}\left\{|\hat{\sigma}_{st} - \sigma_{st}| > \epsilon\right\} \\
\leq \; & \exp\left\{-Cp\min\left(\frac{\epsilon^2}{K_0^2}, \frac{\epsilon}{K_0}\right) + 2\log n\right\}.
\end{aligned}$$

□

8.2 Proofs of Theorem 4.1 and Lemma 4.1

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

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

Thus, we have

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

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

Proof of Lemma 4.1. We consider the following test.

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

The alternative region has decomposition

$$\begin{aligned} H_1 &\subset \Big\{||ZZ^T - Z_0Z_0^T||_F > \sqrt{M}\epsilon, ||ZZ^T + I||_\infty \leq 4K_0\Big\} \\ &\quad \cup \bigcup_{l>1} \Big\{4K_0p^{(l-1)/2} < ||ZZ^T + I||_\infty \leq 4K_0p^{l/2}\Big\} \\ &= \bigcup_{l=0}^{\infty} H_{1l} \end{aligned}$$

Define

$$\begin{aligned} \phi_0 &= \left\{\left\|\frac{1}{p}XX^T - (Z_0Z_0^T + I)\right\|_F > \frac{1}{2}\sqrt{M}\epsilon\right\}, \\ \phi_l &= \left\{\left\|\frac{1}{p}XX^T\right\|_\infty > 2K_0p^{(l-1)/2}\right\}, \quad \text{for each } l. \end{aligned}$$

Then,

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

and

$$\begin{aligned} P_{Z_0}\phi_l &\leq P_{Z_0}\left\{\left\|\frac{1}{p}XX^T - (Z_0Z_0^T + I)\right\| > K_0p^{(l-1)/2}\right\} \\ &\leq \exp\Big(-Cp^{(l+1)/2} + 2\log n\Big), \end{aligned}$$

by Lemma 8.1 and Lemma 8.2. We also have for any $Z \in H_{10}$,

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

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

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

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

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

Thus, the proof is complete. □

8.3 Proof of Theorems 4.2-4.4

Proof of Theorem 4.2. Without loss of generality, we assume n even in the proof. First, notice 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 the pIBP. We choose P to be the stick-breaking representation described in Section 3. That is, under probability P, we first sample $\{p_k\}$ according to (2), 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 stick-breaking representation of pIBP, $\{r_{1k}\}$ and $\{r_{2k}\}$ given $\{p_k\}$ are i.i.d. Bernoulli random variables with parameter $1 - \exp(-\eta\gamma_k)$. Then, z_{ik} are sampled conditioning on (r_{1k}, r_{2k}). When $r_{1k} = 1$,

$z_{ik} = 1$ 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_k determines z_{ik} for $i \in S_2$ in the same way.

We first study $P\Big(||Z - Z_0||_F^2 \Big| \{v_k\}, \alpha\Big)$ for a given α, and then integrate out α with the $\Gamma(1,1)$ distribution. We choose a particular ordering of columns of Z_0. Given the factor decomposition (5), 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

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

We further order the first K_0^* columns of Z_0 so that $\{m_k\}$ decrease. 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.

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

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

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

We study these terms separately. The first two terms are

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

where we have used the inequality $1-q^\beta \geq \beta(1-q)$ for $\beta, q \in (0,1)$ in the last line. The last term is

$$
\begin{aligned}
& P\Big(\sum_{k=K_0+1}^{\infty}\sum_{i=1}^{n} z_{ik} = 0 \Big| \{v_k\}, \alpha\Big) \\
\geq\ & \prod_{k=K_0+1}^{\infty} \exp\Big(-n\gamma_k(1-\eta)\Big) \times P\Big(r_{1k} = r_{2k} = 0, \text{ for } k > K_0 \Big| \{v_k\}, \alpha\Big) \\
\geq\ & \prod_{k=K_0+1}^{\infty} \exp\Big(-n\gamma_k(1-\eta)\Big) \times \prod_{k=K_0+1}^{\infty} \exp(-2\eta\gamma_k) \\
=\ & \prod_{k=K_0+1}^{\infty} (1-p_k)^{n(1-\eta)+2\eta}.
\end{aligned}
$$

The third term is

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

To summarize, we have

$$\begin{aligned}
& P\Big(||Z - Z_0||_F^2 = 0 | \{v_k\}, \alpha\Big) \\
\geq\ & (1-p_1)^{(nK_0^* - M^*)(1-\eta) + 2K_0^*\eta} p_{K_0^*}^{M^*} (1-p_{K_0^*})^{(\eta + n(1-\eta)/2)(K_{01}+K_{02})} p_{K_0}^{K_{01}+K_{02}} \eta^{K_{01}+K_{02}} (1-\eta)^{M^*} \\
& \times \prod_{k=K_0+1}^{\infty} (1-p_k)^{n(1-\eta)+2\eta}.
\end{aligned}$$

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

$$\begin{aligned}
& P\Big(||Z - Z_0||_F^2 = 0 | \{v_k\}, \alpha\Big) \\
\geq\ & 4^{-(nK_0^* - M^*)(1-\eta) - 2K_0^*\eta} \left(1 - (4/3)^{-K_0^*}\right)^{(\eta + n(1-\eta)/2)(K_{01}+K_{02})} \\
& \times 4^{-K_0^* M^*} 4^{-K_0(K_{01}+K_{02})} \eta^{K_{01}+K_{02}} (1-\eta)^{M^*} \\
& \times \prod_{k=K_0+1}^{\infty} \left(1 - (4/3)^{-K_0} \prod_{i=K_0+1}^{k} v_i\right)^{n(1-\eta)+2\eta}.
\end{aligned}$$

The expectation of last term is

$$\begin{aligned}
& \mathbb{E}\left\{ \prod_{k=K_0+1}^{\infty} \left(1 - (4/3)^{-K_0} \prod_{i=K_0+1}^{k} v_i\right)^{n(1-\eta)+2\eta} \right\} \\
\geq\ & \exp\left\{ (n(1-\eta)+2\eta) \sum_{k=K_0+1}^{\infty} \mathbb{E} \log\left(1 - (4/3)^{-K_0} \prod_{i=K_0+1}^{k} v_i\right) \right\} \\
\geq\ & \exp\left\{ -\delta(n(1-\eta)+2\eta)(4/3)^{-K_0} \sum_{k=K_0+1}^{\infty} \mathbb{E}\left(\prod_{i=K_0+1}^{k} v_i\right) \right\} \\
\geq\ & \exp\left\{ -\delta(n(1-\eta)+2\eta)(4/3)^{-K_0} \sum_{k=K_0+1}^{\infty} \left(\frac{\alpha}{\alpha+1}\right)^{k-K_0} \right\} \\
=\ & \exp\Big(-\alpha\delta(4/3)^{-K_0} \big(n(1-\eta)+2\eta\big) \Big),
\end{aligned}$$

where we first apply Jensen's inequality, and then use

$$\log(1-x) \geq -\delta x, \quad \text{for } |x| \leq 1/2,$$

with $\delta > 0$ being a universal constant. The probability of $\mathcal{H}$ is

$$\begin{aligned}
\mathbb{P}(\mathcal{H}) &= \left(\int_{1/4}^{3/4} \frac{t^{\alpha-1}}{\mathrm{B}(\alpha,1)} dt \right)^{K_0} \\
&= M_\alpha^{-K_0},
\end{aligned}$$

where

$$M_\alpha = \frac{\alpha \mathrm{B}(\alpha, 1)}{(3/4)^\alpha - (1/4)^\alpha} > 1.$$

Finally, we have

$$\begin{aligned}
& P\Big(||Z - Z_0||_F^2 = 0\Big) \\
\geq\ & P\Big(||Z - Z_0||_F^2 = 0 \Big| \{v_k\} \in \mathcal{H}, \alpha \in (1/2, 2)\Big) \mathbb{P}\Big(\mathcal{H} \Big| \alpha \in (1/2, 2)\Big) \mathbb{P}\Big(\alpha \in (1/2, 2)\Big) \\
\geq\ & C' \inf_{\alpha \in (1/2,2)} \Big(M_\alpha^{-K_0}\Big) \exp\Big(- 2\delta (4/3)^{-K_0} \big(n(1-\eta) + 2\eta\big)\Big) \\
& \times 4^{-(nK_0^* - M^*)(1-\eta) - 2K_0^*\eta} \left(1 - (4/3)^{-K_0^*}\right)^{(\eta + n(1-\eta)/2)(K_{01} + K_{02})} \\
& \times 4^{-K_0^* M^*} 4^{-K_0(K_{01}+K_{02})} \eta^{K_{01}+K_{02}} (1-\eta)^{M^*} \\
\geq\ & A^{-K_0(K_0 - K_0^* + 1)} \exp\left(- CnK_0^{*2} - Cn(1-\eta) - Cn(1-\eta)\frac{K_0 - K_0^*}{(4/3)^{K_0^*}}\right),
\end{aligned}$$

for some universal constants $A > 1$ and $C > 0$. Observe that the above proof also works by replacing K_0 and K_0^* by $K_0 + \kappa$ and $K_0^* + \kappa$ for any $\kappa > 0$, and 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. □

8.4 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 the pIBP or IBP on $[Z]$. In this case, we use the following theorem instead of Theorem 4.1.

Theorem 8.1. *If* $\Pi\Big((2\sigma_X^4)^{-1}||\sigma_A^2 ZZ^T + \sigma_X^2 I - (\sigma_{A,0}^2 Z_0 Z_0^T + \sigma_{X,0}^2 I)||_F^2 \leq \epsilon^2\Big) \geq \exp\Big(-Cp\epsilon^2\Big)$*, for some* $C > 0$*, and there is a testing function* ϕ*, such that* $P_{Z_0}\phi + \sup_{Z \in U} P_Z(1-\phi) \leq exp\Big(-(C+4)p\epsilon^2\Big)$*, then*

$$P_{Z_0}\Pi\Big(U|X\Big) \longrightarrow 0.$$

Proof. In view of Theorem 2.1 of Ghosal et al. (2000), we only need to lower bound the prior probability of the Kullback-Leibler neighborhood of the truth. According to (1),

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

where the last inequality is because

$$\lambda_{\min} \left(\sigma_A^2 ZZ^T + \sigma_X^2 I \right) \geq \sigma_X^2.$$

In the same way,

$$\begin{aligned}
& \mathrm{Var}_{P_{Z_0}} \left(\log \frac{dN(0, \sigma_{A,0}^2 Z_0 Z_0^T + \sigma_{X,0}^2 I)}{dN(0, \sigma_A^2 ZZ^T + \sigma_X^2 I)} \right) \\
\leq & \frac{1}{2} \left\| \left(\sigma_A^2 ZZ^T + \sigma_X^2 I - (\sigma_{A,0}^2 Z_0 Z_0^T + \sigma_{X,0}^2 I) \right) \left(\sigma_A^2 ZZ^T + \sigma_X^2 I \right)^{-1} \right\|_F^2 \\
\leq & \frac{1}{2\sigma_X^4} \left\| \sigma_A^2 ZZ^T + \sigma_X^2 I - (\sigma_{A,0}^2 Z_0 Z_0^T + \sigma_{X,0}^2 I) \right\|_F^2.
\end{aligned}$$

Therefore, we have

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

Thus, the proof is complete. □

Theorem 8.2. *Assume* $\dfrac{\log p}{n} \to 0$. *Theorem 4.3 and 4.4 still hold if there is a universal constant* $B > 0$, *such that* $\sigma_{A,0}^2 \in (B^{-1}, B)$, $\sigma_{X,0}^2 \in (B^{-1}, B)$ *and* $\inf_{t \in (0,2B)} \pi_A(t) \wedge \inf_{t \in (0,2B)} \pi_X(t) \geq B^{-1}$.

Proof. According to Theorem 8.1 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 prior concentration rate in Section 4.4. in all situations. Using conditioning and the independent structure of the prior, we have

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

because $||Z_0 Z_0^T||_F^2 \le n^2 K_0$ and $||I||_F^2 = n$. The variance part has lower bound

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

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

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

Similarly, when $\dfrac{\epsilon^2}{(1+\epsilon/\sqrt{2n})^2}$ does not go to 0, $\pi_A\left(n^2 K_0 B^2\Big(1+\epsilon/\sqrt{2n}\Big)^2|\sigma_A^2 - \sigma_{A,0}^2|^2 \le \epsilon^2/2\right)$ can be lower bounded by a constant. When it goes to zero, we have

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

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

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

for a constant C_0 only depending on B. Hence, for $\eta = 0$, we have

$$\begin{aligned} & \Pi\Big((2\sigma_X^4)^{-1}||\sigma_A^2 ZZ^T + \sigma_X^2 I - (\sigma_{A,0}^2 Z_0 Z_0^T + \sigma_{X,0}^2 I)||_F^2 \le \epsilon^2\Big) \\ \ge & \exp\Big(-C'\big(\log p + \log n + \log K_0\big)\Big) \times A^{-K_0} \exp\Big(-Cn(K_0^2+1)\Big) \\ \ge & A^{-K_0} \exp\Big(-C_1 n(K_0^2+1)\Big), \end{aligned}$$

for some $C_1 > 0$ because $\dfrac{\log p}{n} \to 0$. This completes Theorem 4.3. For $\eta \in (0,1)$, we have

$$\begin{aligned} & \Pi\Big((2\sigma_X^4)^{-1}||\sigma_A^2 ZZ^T + \sigma_X^2 I - (\sigma_{A,0}^2 Z_0 Z_0^T + \sigma_{X,0}^2 I)||_F^2 \le \epsilon^2\Big) \\ \ge & \exp\Big(-C'\big(\log p + \log n + \log K_0\big)\Big) \times \exp\Big(-2CnK_0^{2(1-\beta)}\Big) \\ \ge & \exp\Big(-C_2 nK_0^{2(1-\beta)}\Big), \end{aligned}$$

for some $C_2 > 0$. Thus, we also prove Theorem 4.4. □

ACKNOWLEDGEMENTS

We thank Yale University Biomedical High Performance Computing Center for computing resources, and NIH grant RR19895 and RR029676-01, which funded the instrumentation.

REFERENCES

Arriola, E., Marchio, C., Tan, D. S., Drury, S. C., Lambros, M. B., Natrajan, R., Rodriguez-Pinilla, S. M., Mackay, A., Tamber, N., Fenwick, K. et al. (2008), "Genomic analysis of the HER2/TOP2A amplicon in breast cancer and breast cancer cell lines," *Laboratory investigation*, 88(5), 491–503.

Barron, A. R. (1988), "The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions," , .

Bell, D., Berchuck, A., Birrer, M., Chien, J., Cramer, D., Dao, F., Dhir, R., DiSaia, P., Gabra, H., Glenn, P. et al. (2011), "Integrated genomic analyses of ovarian carcinoma," *Nature*, 474, 609–615.

Berois, N., Gattolliat, C.-H., Barrios, E., Capandeguy, L., Douc-Rasy, S., Valteau-Couanet, D., Bénard, J., and Osinaga, E. (2013), "GALNT9 Gene Expression Is a Prognostic Marker in Neuroblastoma Patients," *Clinical chemistry*, 59(1), 225–233.

Brennan, C. (2011), "Genomic profiles of glioma," *Current neurology and neuroscience reports*, 11(3), 291–297.

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., Jacobsen, A., Byrne, C. J., Heuer, M. L., Larsson, E. et al. (2012), "The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data," *Cancer discovery*, 2(5), 401–404.

Chen, L., Wang, Y., Ma, D., and Chen, Y. (2006), "Short interfering RNA against the PDCD5 attenuates cell apoptosis and caspase-3 activity induced by Bax overexpression," *Apoptosis*, 11(1), 101–111.

Chène, P. (2003), "Inhibiting the p53–MDM2 interaction: an important target for cancer therapy," *Nature reviews cancer*, 3(2), 102–109.

Diaconis, P., and Freedman, D. (1986), "On the consistency of Bayes estimates," *The Annals of Statistics*, pp. 1–26.

Frullanti, E., Colombo, F., Falvella, F. S., Galvan, A., Noci, S., De Cecco, L., Incarbone, M., Alloisio, M., Santambrogio, L., Nosotti, M. et al. (2012), "Association of lung adenocarcinoma clinical stage with gene expression pattern in noninvolved lung tissue," *International Journal of Cancer*, 131(5), E643–E648.

Gery, S., Sawyers, C. L., Agus, D. B., Said, J. W., Koeffler, H. P. et al. (2002), "TMEFF2 is an androgen-regulated gene exhibiting antiproliferative effects in prostate cancer cells.," *Oncogene*, 21(31), 4739.

Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000), "Convergence rates of posterior distributions," *Annals of Statistics*, 28(2), 500–531.

Griffiths, T. L., and Ghahramani, Z. (2005), Infinite Latent Feature Models and the Indian Buffet Process,, in *In NIPS*, MIT Press, pp. 475–482.

Griffiths, T. L., and Ghahramani, Z. (2011), "The indian buffet process: An introduction and review," *Journal of Machine Learning Research*, 12, 1185–1224.

Hu, G., Chong, R. A., Yang, Q., Wei, Y., Blanco, M. A., Li, F., Reiss, M., Au, J. L.-S., Haffty, B. G., and Kang, Y. (2009), "¡ i¿ MTDH¡/i¿ Activation by 8q22 Genomic Gain Promotes Chemoresistance and Metastasis of Poor-Prognosis Breast Cancer," *Cancer cell*, 15(1), 9–20.

Jariwala, U., Prescott, J., Jia, L., Barski, A., Pregizer, S., Cogan, J. P., Arasheben, A., Tilley, W. D., Scher, H. I., Gerald, W. L. et al. (2007), "Identification of novel androgen receptor target genes in prostate cancer," *Mol Cancer*, 6, 39.

Jordheim, L. P., Nguyen-Dumont, T., Thomas, X., Dumontet, C., and Tavtigian, S. V. (2008), "Differential allelic expression in leukoblast from patients with acute myeloid leukemia suggests genetic regulation of CDA, DCK, NT5C2, NT5C3, and TP53," *Drug Metabolism and Disposition*, 36(12), 2419–2423.

Killian, A., Sarafan-Vasseur, N., Sesboüé, R., Le Pessot, F., Blanchard, F., Lamy, A., Laurent, M., Flaman, J.-M., and Frébourg, T. (2006), "Contribution of the BOP1 gene, located on 8q24, to colorectal tumorigenesis," *Genes, Chromosomes and Cancer*, 45(9), 874–881.

Knowles, D., and Ghahramani, Z. (2011), "Nonparametric bayesian sparse factor models with application to gene expression modeling," *The Annals of Applied Statistics*, 5(2B), 1534–1552.

Lv, Q., Wang, W., Xue, J., Hua, F., Mu, R., Lin, H., Yan, J., Lv, X., Chen, X., and Hu, Z.-W. (2012), "DEDD Interacts with PI3KC3 to Activate Autophagy and Attenuate Epithelial–Mesenchymal Transition in Human Breast Cancer," *Cancer research*, 72(13), 3238–3250.

McLendon, R., Friedman, A., Bigner, D., Van Meir, E. G., Brat, D. J., Mastrogianakis, G. M., Olson, J. J., Mikkelsen, T., Lehman, N., Aldape, K. et al. (2008), "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature*, 455(7216), 1061–1068.

Miller, K. T., Griffiths, T., and Jordan, M. I. (2012), "The phylogenetic indian buffet process: A non-exchangeable nonparametric prior for latent features," *arXiv preprint arXiv:1206.3279*, .

Mo, Q., Wang, S., Seshan, V. E., Olshen, A. B., Schultz, N., Sander, C., Powers, R. S., Ladanyi, M., and Shen, R. (2013), "Pattern discovery and cancer gene identification in integrated cancer genomic data," *Proceedings of the National Academy of Sciences*, 110(11), 4245–4250.

Muzny, D. M., Bainbridge, M. N., Chang, K., Dinh, H. H., Drummond, J. A., Fowler, G., Kovar, C. L., Lewis, L. R., Morgan, M. B., Newsham, I. F. et al. (2012), "Comprehensive molecular characterization of human colon and rectal cancer," *Nature*, 487, 330–337.

Neilsen, P. M., Noll, J. E., Suetani, R. J., Schulz, R. B., Al-Ejeh, F., Evdokiou, A., Lane, D. P., Callen, D. F. et al. (2011), "Mutant p53 uses p63 as a molecular chaperone to alter gene expression and induce a pro-invasive secretome," *Oncotarget*, 2(12), 1203.

Nik-Zainal, S., Van Loo, P., Wedge, D. C., Alexandrov, L. B., Greenman, C. D., Lau, K. W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M. et al. (2012), "The life history of 21 breast cancers," *Cell*, 149(5), 994–1997.

Nord, H., Hartmann, C., Andersson, R., Menzel, U., Pfeifer, S., Piotrowski, A., Bogdan, A., Kloc, W., Sandgren, J., Olofsson, T. et al. (2009), "Characterization of novel and complex genomic aberrations in glioblastoma using a 32K BAC array," *Neuro-oncology*, 11(6), 803–818.

Ramakrishna, M., Williams, L. H., Boyle, S. E., Bearfoot, J. L., Sridhar, A., Speed, T. P., Gorringe, K. L., and Campbell, I. G. (2010), "Identification of candidate growth promoting genes in ovarian cancer through integrated copy number and expression analysis," *PloS one*, 5(4), e9983.

Rousseau, J., Têtu, B., Caron, D., Malenfant, P., Cattaruzzi, P., Audette, M., Doillon, C., Tremblay, J. P., and Guérette, B. (2002), "RCAS1 is associated with ductal breast cancer progression," *Biochemical and biophysical research communications*, 293(5), 1544–1549.

Samarakoon, R., Higgins, C. E., Higgins, S. P., and Higgins, P. J. (2009), "TGF-1-Induced Expression of the Poor Prognosis SERPINE1/PAI-1 Gene Requires EGFR Signaling: A New Target for Anti-EGFR Therapy," *Journal of oncology*, 2009.

Schwartz, L. (1965), "On bayes procedures," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 4(1), 10–26.

Tasaki, M., Shimada, K., Kimura, H., Tsujikawa, K., and Konishi, N. (2011), "ALKBH3, a human AlkB homologue, contributes to cell survival in human non-small-cell lung cancer," *British journal of cancer*, 104(4), 700–706.

TCGA (2012), "Comprehensive molecular portraits of human breast tumours," *Nature*, 490, 61–70.

Teh, Y. W., Görür, D., and Ghahramani, Z. (2007), Stick-breaking construction for the Indian buffet process,, in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, Vol. 11.

Thomadaki, H., Talieri, M., Scorilas, A. et al. (2007), "Prognostic value of the apoptosis related genes BCL2 and BCL2L12 in breast cancer.," *Cancer letters*, 247(1), 48.

Van De Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J. et al. (2002), "A gene-expression signature as a predictor of survival in breast cancer," *New England Journal of Medicine*, 347(25), 1999–2009.

Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., Miller, C. R., Ding, L., Golub, T., Mesirov, J. P. et al. (2010), "An integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR and NF1," *Cancer cell*, 17(1), 98.

Vershynin, R. (2010), "Introduction to the non-asymptotic analysis of random matrices," *arXiv preprint arXiv:1011.3027*, .

West, R. B., Corless, C. L., Chen, X., Rubin, B. P., Subramanian, S., Montgomery, K., Zhu, S., Ball, C. A., Nielsen, T. O., Patel, R. et al. (2004), "The Novel Marker, DOG1, Is Expressed Ubiquitously in Gastrointestinal Stromal Tumors Irrespective of KIT or PDGFRA Mutation Status," *The American journal of pathology*, 165(1), 107–113.

Wrzeszczynski, K. O., Varadan, V., Byrnes, J., Lum, E., Kamalakaran, S., Levine, D. A., Dimitrova, N., Zhang, M. Q., and Lucito, R. (2011), "Identification of tumor suppressors and oncogenes from genomic and epigenetic features in ovarian cancer," *PLoS One*, 6(12), e28503.

Wu, J., Lei, G., Mei, M., Tang, Y., Li, H., Wu, J., Lei, G., Mei, M., Tang, Y., and Li, H. (2010), "A Novel C53/LZAP-interacting Protein Regulates Stability of C53/LZAP and DDRGK Domain-containing Protein 1 (DDRGK1) and Modulates NF-κB Signaling," *Journal of Biological Chemistry*, 285(20), 15126–15136.

Yates, L. R., and Campbell, P. J. (2012), "Evolution of the cancer genome," *Nature Reviews Genetics*, 13, 795–806.

Zhang, D., Jiang, P., Xu, Q., Zhang, X., Zhang, D., Jiang, P., Xu, Q., and Zhang, X. (2011), "Arginine and glutamate-rich 1 (ARGLU1) interacts with mediator subunit 1 (MED1) and is required for estrogen receptor-mediated gene transcription and breast cancer cell growth," *Journal of Biological Chemistry*, 286(20), 17746–17754.

List of Figures

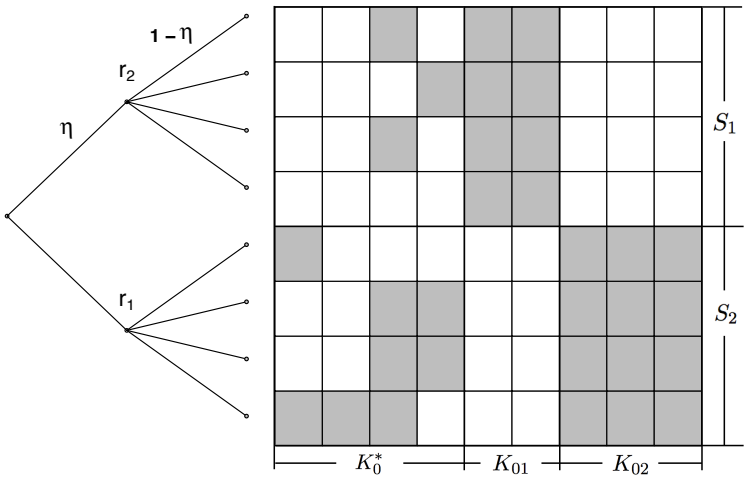

Figure 1: An illustration of the two group tree and the factor decomposition.

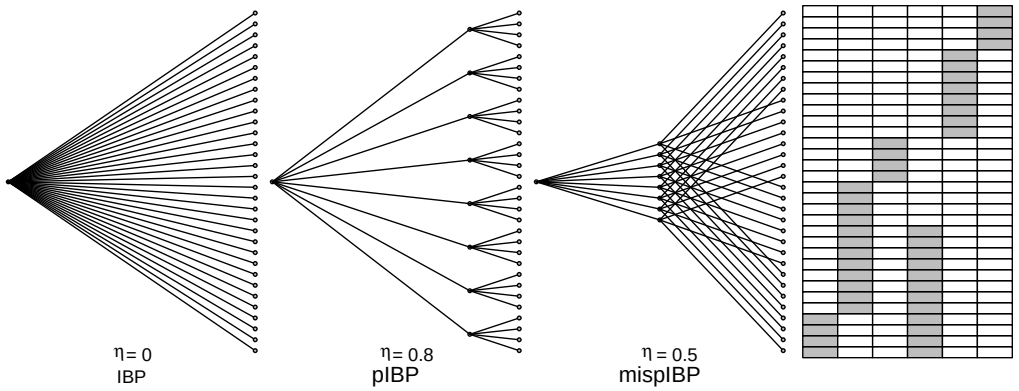

Figure 2: The illustration of the IBP, pIBP with an appropriate tree structure and pIBP with a mis-specified tree structure and the latent factor matrix Z_0 used in the first simulation.

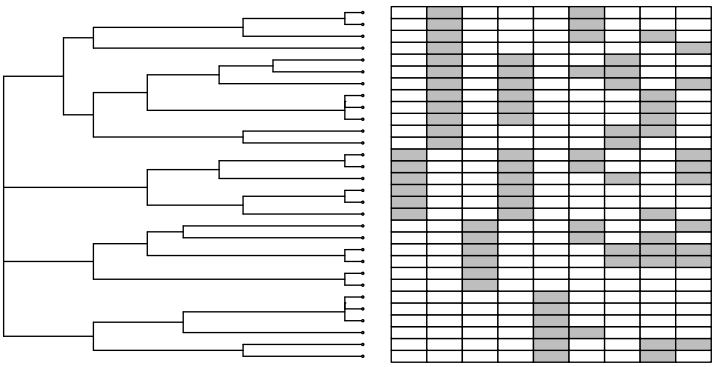

Figure 3: The illustration of the latent factor matrix Z_0 and tree prior constructed from the hierarchical clustering analysis of Z_0 in the second simulation.

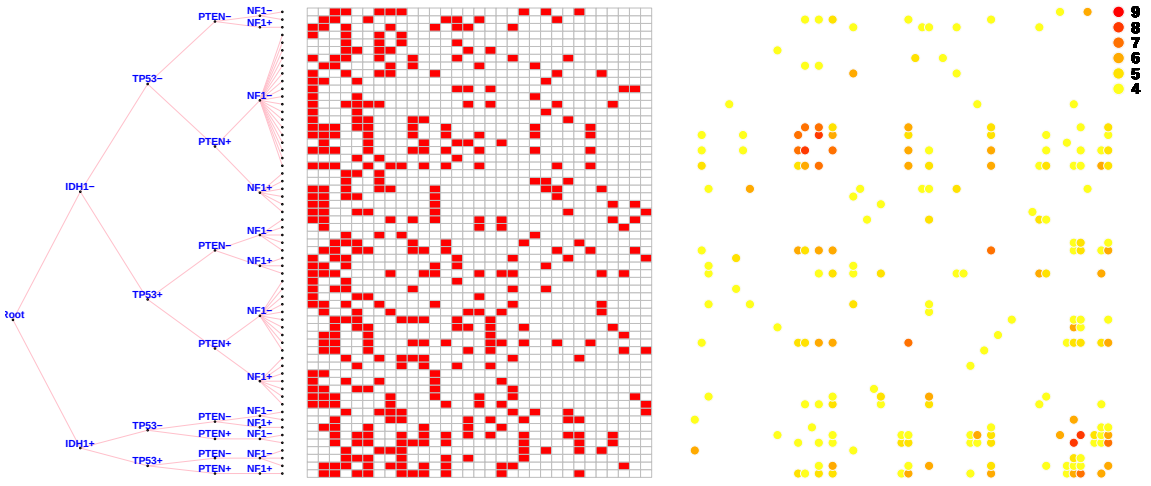

Figure 4: A graph showing the logic tree prior (left), the inferred latent factor matrix Z (middle) and the feature similarity matrix ZZ^T(right) for TCGA GBM dataset.

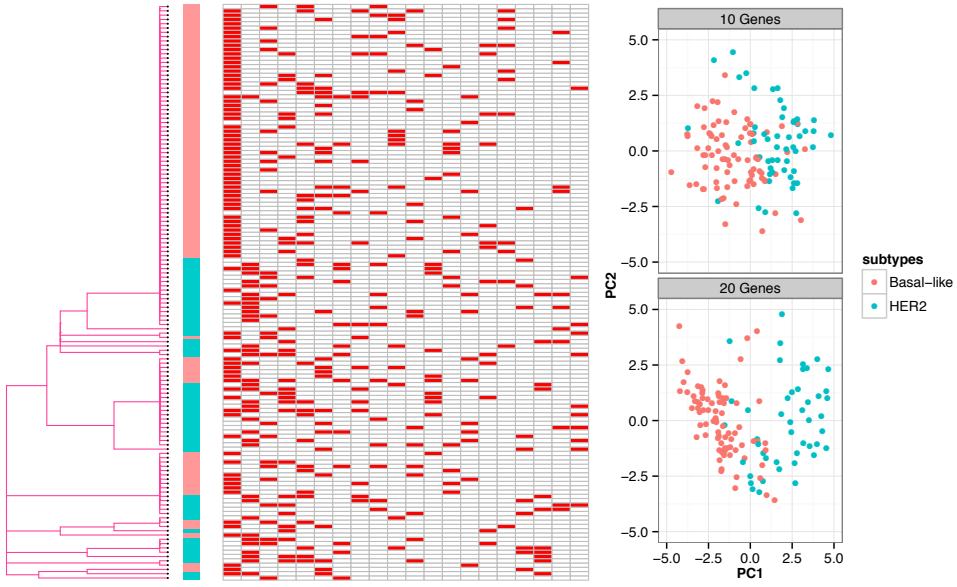

Figure 5: A graph showing the dendrogram tree prior (left), the inferred latent factor matrix Z (middle, only first 20 columns shown) and PCA analysis of Basal-like (Red) and HER2 (Green) based on genes with top loading on latent factors (topright, with a set of 10 genes from first factor; bottomright, with a set of 20 genes from first two factors) for TCGA BRCA dataset.

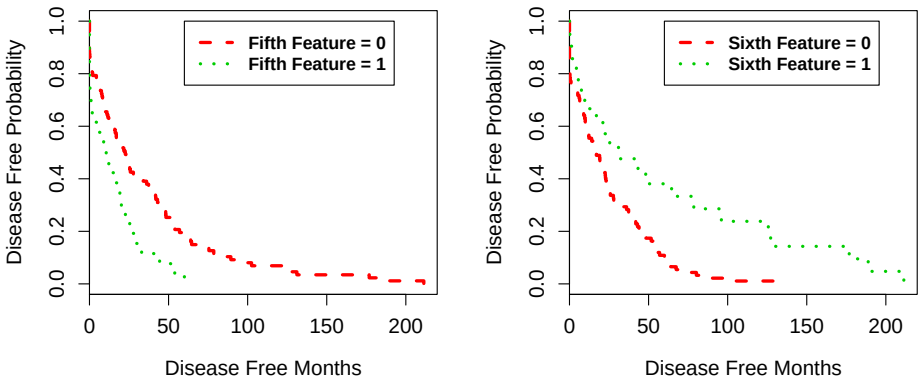

Figure 6: A Kaplan - Meier plot for groups with different status of the fifth and sixth feature inferred from TCGA BRCA dataset.

List of Tables

Table 1: Simulation results: comparisons of IBP, pIBP with the appropriate tree prior and pIBP with the mis-specified tree prior (mispIBP) based on 40 independent replicates. The performance is measured by estimation errors in terms of Frobenius norm of the feature similarity matrix $\sqrt{n^{-1}||ZZ^T - Z_0Z_0^T||_F^2}$ (F-norm), and the number of estimated latent factors $\hat{K}$. Numbers in parentheses are the standard deviations across the 40 independent replicates.

(n,p)	IBP		pIBP		mispIBP	
	F-norm	$\hat{K}$	F-norm	$\hat{K}$	F-norm	$\hat{K}$
(192,20)	18.9 (15.2)	8.1 (3.8)	6.8 (2.3)	6.7 (1.1)	16.2 (9.1)	6.6 (0.8)
(288,20)	20.3 (8.9)	7 (1.9)	10 (2.2)	7 (0.9)	19.3 (14.6)	7 (0.9)
(384,20)	27.8 (7.4)	7.5 (1.8)	16 (7.7)	7.9 (2.4)	32.3 (5.8)	7.8 (1.3)
(192,30)	9.5 (6.9)	6.6 (0.8)	4.9 (3)	6.1 (0.3)	14.4 (15.3)	6.8 (1.6)
(288,30)	14.2 (5.2)	6.6 (0.5)	7.9 (6.1)	6.6 (1.4)	13.2 (12.5)	6.4 (0.6)
(384,30)	14.5 (8.2)	6.7 (0.9)	8 (4.8)	6.4 (0.7)	13.9 (9.7)	6.7 (0.8)
(192,100)	3.8 (2.3)	5.9 (0.6)	4 (2.2)	5.8 (0.6)	3.8 (2.2)	5.9 (0.6)
(288,100)	5.5 (2.3)	5.8 (0.5)	5.2 (2)	5.8 (0.6)	5.3 (2.1)	5.8 (0.5)
(384,100)	6 (3.4)	6 (0.6)	5.5 (3.9)	6.2 (0.9)	5.7 (3.4)	6 (0.8)
(192,200)	3.8 (1.8)	5.8 (0.6)	3.8 (1.9)	5.5 (1.1)	3.8 (1.9)	5.5 (1.1)
(288,200)	4.8 (2.3)	5.7 (0.5)	4.8 (2.3)	5.7 (0.5)	4.9 (2.4)	5.7 (0.5)
(384,200)	5 (2.4)	5.6 (0.6)	4.7 (2.6)	5.6 (0.5)	4.6 (2.5)	5.7 (0.6)

In the above models, $\sigma_{A,0}^2 = 1$, $\sigma_{X,0}^2$ =0.5, $K_0 = 6$, results are based on 1000 MCMC steps.

Table 2: Simulation results: comparisons of IBP and pIBP with the tree prior from the dendrogram of a hierarchical clustering on Z_0 based on 40 independent replicates.

(n, p)	IBP		pIBP		mispIBP	
	F-norm	$\hat{K}$	F-norm	$\hat{K}$	F-norm	$\hat{K}$
(120, 15)	28.5 (6)	22.5 (1.6)	11.4 (6.4)	17 (3.9)	31.1 (10.5)	23.6 (3.4)
(180, 15)	30.4 (3.9)	21.5 (1.4)	11.9 (4.7)	15.5 (2.9)	31.2 (7.1)	23.1 (3.1)
(240, 15)	35 (7.2)	18.5 (4.9)	13.4 (2.3)	17.8 (2.5)	32.6 (4.3)	24.6 (2)
(120, 30)	11.8 (7.7)	11.9 (3.6)	7 (2.3)	11.7 (2.5)	8.1 (3.5)	11.6 (1.5)
(180, 30)	13.9 (6.9)	12.3 (3)	9.2 (2.9)	13.3 (2.7)	12.1 (3.3)	12.4 (1.8)
(240, 30)	15.9 (10.4)	12.2 (3.3)	10.7 (3)	13.2 (2.2)	18.2 (8.4)	11.1 (1.4)
(120, 60)	7.3 (2.8)	11.2 (1.5)	6.7 (2.3)	10.6 (1.5)	7.6 (2.5)	10.6 (1.5)
(180, 60)	9.6 (2.5)	11.7 (2.2)	8.1 (2.5)	11.1 (2.3)	9.4 (3.9)	10.8 (1.2)
(240, 60)	9.4 (3.2)	11.5 (2.4)	9.3 (2.2)	10.8 (1.6)	11.7 (4.2)	11.3 (1.7)

In the above models, $\sigma^2_{A,0} = 1$, $\sigma^2_{X,0} = 0.5$, $K_0 = 9$, results are based on 1000 MCMC steps.