HHS Public Access

Author manuscript

Bayesian Anal. Author manuscript; available in PMC 2017 June 01.

Published in final edited form as:

Bayesian Anal. 2016 June ; 11(2): 477–497. doi:10.1214/15-BA958.

Posterior Contraction Rates of the Phylogenetic Indian Buffet Processes

Mengjie Chen[1], **Chao Gao**[2], and **Hongyu Zhao**[1]

[1] University of North Carolina, Chapel Hill

[2] Yale University

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 example that can be used to define a prior distribution on infinite binary features, where the exchangeability among subjects is assumed. The 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 the theoretical properties of IBP and pIBP under a binary factor model. We establish the posterior contraction rates for both IBP and pIBP and substantiate the theoretical results through simulation studies. This is the first work addressing the frequentist property of the posterior behaviors of IBP and pIBP. We also demonstrated its practical usefulness by applying pIBP prior to a real data example arising in the field of cancer genomics where the exchangeability among subjects is violated.

Keywords

Bayesian Nonparametrics; Indian Buffet Process; Latent Factor Analysis; Cancer Genomics

1 Introduction

Recently nonparametric Bayesian approaches have become popular methods in machine learning and other fields to learn structural information from data. By expressing prior distributions as general stochastic processes, nonparametric Bayesian methods provide flexible ways 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 is presented in the form of a binary matrix with a finite number of rows and an infinite number of columns [18, 22]. 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 cancer patients with different subtypes. In these cases, treating all subjects exchangeable using IBP is not appropriate. As an alternative, the phylogenetic Indian buffet process (pIBP) [26] provides a flexible framework to incorporate prior structural information among 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 and allowance of a potentially infinite number of latent factors. In addition, pIBP provides an effective approach to incorporate useful information on the relationship among subjects without losing computational tractability.

Despite many successful applications of IBP and its variants in many areas [19], as far as we know, there has not been any theoretical investigation of their posterior behaviors. Suppose there is a true data-generating process, do the posterior distributions of IBP and pIBP concentrate on the truth? In the parametric setting where the number of parameters is fixed, the posterior distribution is well behaved according to the classical Bernstein-von Mises theorem [23]. However, when the prior charges a diverging or an infinite number of parameters, whether the posterior distribution still possesses such convergence properties is no longer guaranteed. IBP prior and pIBP prior belong to the second situation because they are stochastic processes on infinite binary matrices. Besides the issue of posterior convergence, we are also interested in the question whether the extra information in pIBP prior would lead to better posterior behavior than that of IBP prior.

In this paper, we study the theoretical properties of IBP and pIBP under a binary factor model. Posterior contraction rates are derived for both priors under various settings. By imposing a group structure on the true binary factor matrix, pIBP is proved to have faster convergence rates than IBP whenever the group structure is well-specified by the phylogenetic tree. Even when the group structure is mis-specified by pIBP, it still has the same convergence rate as that of IBP. To the best of our knowledge, this is the first work addressing the frequentist property of the posterior behaviors of both IBP and pIBP.

We further substantiated the theoretical results through simulation studies. Our simulations show that pIBP is an attractive alternative to IBP when subjects can be related through a tree structure based on some prior information. Moreover, even when the tree structure is mis-specified in pIBP prior, the posterior behavior is still comparable to that of IBP prior, suggesting a robust property of pIBP. We further apply pIBP to analyze cancer genomics data to demonstrate its practical usefulness.

We organize the rest of the paper as follows. Section 2 introduces a binary factor model, which is the probabilistic setting of the paper. 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. Sections 6 presents the analysis of a TCGA data set using pIBP. Section 7 discusses related work on factor models and an extension of our theoretical results. Proofs for theoretical results are collected in the supplementary materials.

2 Problem Setting

2.1 Notation

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 $a_n \leq Cb_n$ for all n. For a matrix $A = (a_{ij})_{m\times n}$, denote its matrix Frobenius norm by $\|A\|_F = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2\right)^{1/2}$. For a set S, denote its cardinality by $|S|$. The symbol Π stands for the prior probability distribution associated with the mixture of IBP or pIBP defined in Section 3.4, and $\Pi(\cdot|X)$ is the corresponding posterior distribution.

2.2 Binary Factor model

Let $X = (x_{ij})_{n\times p}$ denote the observed data matrix, where each of the n rows represents one individual and each of the p columns represents one measurement. We hypothesize that the measurement profiles can be characterized by latent factors. We model the effects of these latent factors Z on X 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 a loading matrix. The status of z_{ik}, which takes a value of 1 or 0, indicates the presence or the absence of the kth factor in the ith individual. The value of a_{kj} weighs the contribution to the jth measurement from the kth factor. We assume that each entry of $E = (e_{ij})_{n\times p}$ follows $N\left(0, \sigma_X^2\right)$ independently. Let each entry of A follow $N\left(0, \sigma_A^2\right)$ independently, and A is independent of E. Conditioning on A, $(X \mid A)$ follows a matrix normal distribution with mean ZA. Integrating out A with respect to its distribution, each column of X follows

$$(x_{1j}, \ldots, x_{nj})^{\mathrm{T}} \sim N\left(0, \sigma_A^2 ZZ^T + \sigma_X^2 I\right), \quad (1)$$

independently for $j = 1, \ldots, p$. Formula (1) shows the covariance structure across individuals imposed by the binary factor model. From this representation, it is easy to see that the matrix ZZ^{T} and the variance components σ_A^2 and σ_X^2 uniquely determine the data generating process.

2.3 Feature Similarity Matrix ZZ^{T}

We name ZZ^{T} the feature similarity matrix because of its important statistical meaning as reflected in (1). An identifiability issue is that the distribution of (1) will not change if one reorder the columns of the factor matrix Z. Thus, Z is not identifiable in the model. However, the feature similarity matrix ZZ^{T}, according to (1), is identifiable. We denote each element of this matrix by $ZZ^{\mathrm{T}} = (\xi_{ij})_{n\times n}$. Each row/column of this matrix ZZ^{T} describes the feature similarity between a particular individual and the other $n - 1$ individuals. Note that

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

Thus, the diagonal element ξ_{ii} denotes the number of factors possessed by the ith individual, and the off-diagonal entry ξ_{ij} is the number of the factors shared between the ith and jth individuals. In short, the feature similarity matrix ZZ^{T} characterizes the latent feature sharing structure among samples. For the ith individual, we define $d_i = \Sigma_{j\neq i}\,\xi_{ij}$ 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 Bu et Process Prior

3.1 A Bayesian Framework

To pursue a full Bayesian approach, we put a prior distribution on the triple $\left(Z,\sigma_A^2,\sigma_X^2\right)$. The choice of the prior on $\left(\sigma_A^2,\sigma_X^2\right)$ is not essential, because for asymptotic purpose (when n and p are large), the prior e ect on the parametric part $\left(\sigma_A^2,\sigma_X^2\right)$ 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 diverges, Bayesian method is no longer guaranteed to be consistent [11]. Thus, the choice of the prior on Z is important. According to 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 a prior on Z, we specify a 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 [18], and its tree-structured generalization, the phylogenetic Indian buffet process proposed by [26]. Both are priors on sparse infinite binary matrices.

3.2 Indian Buffet Process

We describe the Indian buffet process (IBP) on $[Z]$ by its stick-breaking representation derived in [34]. Given some $\alpha > 0$, first draw $v_k \sim \text{Beta}(\alpha, 1)$ ($k = 1, 2, \ldots$ independently and identically distributed. Then, p_k is

$$p_k=\prod_{i=1}^{k} v_i \quad (k=1,2,\ldots). \quad (2)$$

Given $\{p_k\}$, z_{ik} is drawn independently from a Bernoulli distribution with parameter p_k for $i = 1, \ldots, n$ and $k = 1, 2, \ldots$. The final matrix Z drawn in this way has dimension $n \times K^{+}$, where K^{+} is the number of nonzero columns. According to [18], K^{+} follows a Poisson distribution with mean $\alpha\sum_{k=1}^{n} k^{-1}$. Thus, it is finite with probability 1. The IBP prior on $[Z]$

is the image measure induced by the equivalence map $Z \mapsto [Z]$. A larger α indicates a larger K^+ in the prior modeling.

3.3 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 kth column of Z are not independent in pIBP. Their dependency structure is captured by a stochastic process on a rooted tree similar to the models used in phylogenetics [26]. 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 kth column of Z. First, assign 0 to the root of the tree. Along any path from the root to a leaf, let the value of any node 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. pIBP prior is defined to be the image measure on $[Z]$.

3.4 A Hyperprior on α

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 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 after α is integrated out.

4 Posterior Contraction Rates of IBP and pIBP

4.1 Convergence of the Feature Similarity Matrix

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.

We define the triple $\left(Z_0, \sigma^2_{A,0}, \sigma^2_{X,0}\right)$ 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. Denote the data generating process of (1) by P_Z, and let E_Z be the associated expectation (and similarly define P_{Z0} and E_{Z0}). The generalization to the case where $\left(\sigma^2_A, \sigma^2_X\right)$ is unknown is covered in the supplementary materials. Let Π be the mixture of IBP or pIBP prior on $[Z]$. Note that the matrix ZZ^{T} does not depend on the order of columns of Z, and thus we have $ZZ^{\mathrm{T}} = [Z][Z]^{\mathrm{T}}$. We consider the posterior convergence in the sense of

$$E_{Z_0}\left[\Pi\left(\left\|ZZ^T - Z_0 Z_0^{\mathrm{T}}\right\|_F^2 \leq M\epsilon_{n,p}^2 \middle| X\right)\right] \geq 1 - \delta_{n,p}, \quad (3)$$

for some sequences $\epsilon_{n,p}$, $\delta_{n,p}$ and constant $M > 0$. When $\delta_{n,p} \to 0$, this is called posterior contraction of feature similarity matrix with rate $\epsilon^2_{n,p}$ under the squared Frobenius loss. We

choose to study the posterior contraction in terms of the feature similarity matrix ZZ^{T} because of both the identifiability issue and statistical interpretation described in Section 2.3.

4.2 A General Method for Discrete Priors

The theory of Bayesian posterior consistency was first studied by [32]. She proposed a Kullback-Leibler property of the prior and a testing argument to prove weak consistency in the parametric case. The first nonparametric posterior consistency result was obtained by [2], where the idea of testing on the essential support of the prior is used. Later, the same argument was modified to achieve rate of contraction by [17]. In the current setting of binary factor model, we propose the following general method to prove posterior rate of contraction for priors supported on a discrete set.

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

$$E_{Z_0}\left[\Pi\left(U|X\right)\right] \leq E_{Z_0}\left(\phi\right) + \frac{1}{\Pi\left(\left\|ZZ^{\mathrm{T}} - Z_0Z_0^{\mathrm{T}}\right\|_F^2 = 0\right)} \sup_{Z\in U} E_Z\left(1-\phi\right). \quad (4)$$

The theorem can be viewed as a discrete version of the Schwartz theorem [32]. We take advantage of the discrete nature of the problem, thus avoiding calculating the prior mass of the Kullback-Leibler neighborhood of P_{Z0}. We specify U to be

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

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 $\Pi\left(\left\|ZZ^{\mathrm{T}} - Z_0Z_0^{\mathrm{T}}\right\|_F^2 = 0\right)$ and constructing a testing function for $H_0 : Z = Z_0$ and $H_1 : Z \in U$ with appropriate type 1 and type 2 error bounds. The existence of such testing function 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 $E_{Z_0}\left(\phi\right) + \sup_{\left\{\left\|ZZ^{\mathrm{T}} - Z_0Z_0^{\mathrm{T}}\right\|_F^2 > M\epsilon_{n,p}^2\right\}} E_Z\left(1-\phi\right)$ *is upper bounded by*

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

for some universal constant $C > 0$ *and M introduced in (3).*

Therefore, it is sufficient to lower bound the prior mass $\Pi\left(\left\|ZZ^{\mathrm{T}} - Z_0Z_0^{\mathrm{T}}\right\|_F^2 = 0\right)$ to obtain (3).

4.3 Two-Group Tree and Factor Decomposition

Before studying the prior mass lower bound of IBP and pIBP, we need to specify a non-exchangeable structure among the subjects. To demonstrate the power of pIBP to model non-exchangeability, we study a special but representative tree structure, the two-group tree. Let n individuals be labeled by $\{1, \ldots, n\}$. Without loss of generality, we assume n is even. Let $\{1, \ldots, n\} = S_1 \cup S_2$, where $S_1 = \{1, \ldots, n/2\}$ and $S_2 = \{n/2 + 1, \ldots, n\}$. 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 ith 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.

Our theory covers three cases. The first case is 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 misspecified. 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 factor decomposition

$$K_0 = K_{01} + K_{02} + K_0^*, \quad (5)$$

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. 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 true Z_0 (see Figure 6). Generally speaking, the smaller K_0^* is, the better the group structure is specified by Π.

4.4 Prior Mass

Under the two-group structure defined above, we obtain the following prior mass lower bound.

Theorem 4.2—*For any constant $\eta \in [0, 1)$, there exists some constant $C > 0$ such that the prior mass $\Pi\left(\|ZZ^{\mathrm{T}} - Z_0 Z_0^{\mathrm{T}}\|_F^2 = 0\right)$ can be lower bounded by*

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

for any $\kappa \geq 0$.

Theorem 4.2 provides an explicit characterization of the prior mass lower bound as a function of K_0^*. For a larger K_0^*, the prior mass will be at a smaller order due to an increased level of misspecification. The prior mass lower bound directly determines the posterior contraction rate according to Theorem 4.1 and Lemma 4.1. In the following, we consider $\eta = 0$ and $\eta \in (0, 1)$, separately.

When $\eta = 0$, 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

$$\Pi\left(\|ZZ^{\mathrm{T}} - Z_0 Z_0^{\mathrm{T}}\|_F^2 = 0\right) \geq exp\left(-C_1 n K_0^2\right). \quad (6)$$

The prior mass lower bound for IBP in (6) is the benchmark for us to compare IBP with pIBP in various situations.

When $\eta \in (0, 1)$, 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 [26]. We say the group structure is e ectively 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 $\kappa \geq 0$. That is, for n su ciently large $n \gtrsim K_0^{2\beta}$, we have

$$\Pi\left(\|ZZ^{\mathrm{T}} - Z_0 Z_0^{\mathrm{T}}\|_F^2 = 0\right) \geq exp\left(-C_2 n \min_{k \geq K_0^*}\left(k^2 \vee \frac{K_0}{(4/3)^k}\right)\right), \quad (7)$$

which is lower bounded by

$$exp\left(-C_2' n K_0^{2(1-\beta)}\right).$$

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 $K_0^* \lesssim K_0^{1-\beta}$.

On the other hand, the group structure is mis-specified if $K_0^* = K_0$. In this case, we reduce to (6), so that

$$\Pi\left(\|ZZ^{\mathrm{T}} - Z_0 Z_0^{\mathrm{T}}\|_F^2 = 0\right) \geq exp\left(-C_3 n K_0^2\right).$$

Thus, a mis-specified tree structure does not compromise the results, compared to a default tree structure of IBP. One may wonder whether this is due to a possibly loose bound in Theorem 4.2. By scrutinizing the proof, we found that the slack is at most at a constant level independent of (n, K_0, K_0^*). Thus, the prior mass lower bounds of pIBP with a mis-specified tree and of IBP are essentially the same.

4.5 Posterior Contraction Rates

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

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

$$E_{Z_0}\left[\Pi\left(\|ZZ^{\mathrm{T}}-Z_0Z_0^{\mathrm{T}}\|_F^2 \le M\frac{K_0^4 n^3}{p}|X\right)\right] \ge 1-exp\left(-C'nK_0^2\right),$$

as long as $nK_0^2/p=o(1)$.

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* $K_0^* \lesssim K_0^{1-\beta}$ *and* $K^{2\beta} \lesssim n$ *for* $\beta \in (0, 1)$, *for the binary factor model, there exist* $M > 0$ *and* $C' > 0$, *such that*

$$E_{Z_0}\left[\Pi\left(\|ZZ^{\mathrm{T}}-Z_0Z_0^{\mathrm{T}}\|_F^2 \le M\frac{K_0^{4-2\beta} n^3}{p}|X\right)\right] \ge 1-exp\left(-C'nK_0^{2(1-\beta)}\right),$$

as long as $nK_0^{2(1-\beta)}/p=o(1)$.

The above two theorems establish rates of contraction for the posterior distributions of IBP and pIBP. The posterior probabilities on the neighborhood of the truth can be arbitrarily close to 1 in expectation under the true model for sufficiently large n, p and K_0. The contraction rate is faster for larger p and smaller n, because more variables are helpful to identify the feature similarity of a group of individuals.

Compared with the rate of IBP in Theorem 4.3, when the tree structure is effectively specified, the upper bound of the rate of pIBP in Theorem 4.4 is faster by a factor of $K_0^{2\beta}$. Such difference is significant if the number of features K_0 is large. Moreover, Theorem 4.3 also suggests that even when the tree structure of pIBP is mis-specified, the rate of contraction 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 can also be obtained under a more complicated structural assumption using the same method.

5 Simulation Studies

In this section, we perform simulations to evaluate the performances of IBP and pIBP. We implemented the Markov chain Monte Carlo algorithm proposed in [26] to perform posterior inference of the feature similarity matrix ZZ^{T}. In the algorithm, the sampling process on the tree structure is expressed as a graphical model, where the prior probabilities can be calculated effciently by a sum-product algorithm. All the parameters σ_A, σ_X, α and $\{p_k\}$ (marginal probabilities of a latent feature equaling 1) are sampled as part of the overall Markov chain Monte Carlo procedure.

In the first simulation, we evaluated the performance of IBP, pIBP with a correctly specified 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 illustrated by Figure 2. Twelve models

presented in Table 1 are considered. For each model, we generated an $n \times p$ matrix $X = Z_0A + E$ with $(\sigma_{A,0}, \sigma_{X,0}) = (1,0{\cdot}5)$. For IBP, we let $\eta = 0$ so that pIBP is equivalent to IBP. For pIBP, we let $\eta = 0{\cdot}8$ and a proper tree structure is given. For the mispIBP, we let $\eta = 0{\cdot}5$ 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 normalized Frobenius norm of the feature similarity matrix $n^{-1/2}\|ZZ^{\mathrm{T}} - Z_0Z_0^{\mathrm{T}}\|_F$. We further evaluated the latent structure recovery by the number of estimated latent factors. We observed that both IBP and pIBP overestimates the number of latent factors because of the presence of many factors with only a few samples. This is similar to what was proved for Dirichlet and Pitman-Yor processes where the posterior is inconsistent for estimating the number of clusters [25]. Therefore, we reported a truncated estimator of the number of latent factors counting only those factors shared by at least 5 samples.

The algorithm of [26] is implemented for 1000 MCMC steps. We observe that it guarantees convergence in the problem sizes that are considered in this simulation.

Generally, the 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. 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 important role. Inferences using different priors lead to similar results.

In the second simulation, we used the similarity data to construct 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, we first apply a hierarchical cluster analysis with complete linkage on the rows of Z_0 and then use its output dendrogram as the tree in the pIBP prior (see Figure 3). In our analysis, we constructed our prior based on the true knowledge of Z_0 in order to investigate whether the correct structural information will improve the performance through pIBP priors. In practice, such trees need to be constructed from external sources. 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. Similar as the previous simulation, we evaluated the performance by $n^{-1/2}\|ZZ^{\mathrm{T}} - Z_0Z_0^{\mathrm{T}}\|_F$ and the truncated number of estimated latent features (Table 2). When p is small, pIBP outperforms IBP in all cases. When p is adequately large ($p = 60$ in this setting), the inference is less influenced by the prior information.

6 Applications of pIBP in the Integrative Cancer Genomics Analysis

Cancer research has been revolutionized by recent advances in high through-put technologies. Diverse types of genomics data, e.g., DNA, RNA, and epigenetic, have been profiled for different tumor types [28, 27, 3, 33]. These data have revealed that substantial heterogeneities exist across tumor types, across individuals within the same tumor types and even within an individual tumor. However, the tumor heterogeneity at somatic level has not been explicitly explored in the integrative analysis.

Here we propose to use binary factor model to integrate somatic mutation and gene expression data based on 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 build a pIBP prior based on somatic mutation data then specify it on the latent factors of gene expression data. Using this approach, we can investigate the gene expression data by taking into account the heterogeneities across cancer patients at somatic level.

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 lth gene on the ith 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 containing mutations at the lth gene.

As for using a tree structure to express the relationships of individuals using the somatic mutation data, we propose to construct either logic tree or dendrogram 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. 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.

We analyzed the TCGA BRCA Level 3 dataset generated by [33] (downloaded from cBio [8]) using the dendrogram 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 [33]. 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 Markov chains. No substantial difference was observed across runs and we chose the one with largest posterior probability as the final result. Figure 4 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 [21]; MTBP inhibits cancer metastasis by interacting with MDM2 [10]; DEDD interacts with PI3KC3 to activate autophagy and attenuate epithelialmesenchymal transition in cancer [24]; and HSF1 has been proposed as a predictor of survival in breast cancer [35]. EIF2C2, PUF60 and PHF20L1 have been reported as prognostic markers in ovarian cancer [30, 38], which is consistent with the recent discovery that basal-like breast tumours with high-grade serous ovarian tumours share many molecular commonalities [33]. 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 [1]; MED1 is required for estrogen receptor-mediated gene transcription and breast cancer cell growth [39]. As revealed by principal component analysis based on gene expression (Figure 4), these genes weighing high on first two latent features have discriminating power on Basal-like and HER2 samples.

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 5 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 [20]; EBAG9 (RCAS1) is associated with ductal breast cancer progression [31]. The other 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 o ers 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.

7 Discussion

7.1 Related Work on Factor Models

This paper attempts to provide a theoretical foundation for the widely used IBP and pIBP priors. We illustrate the performance of the priors through a simple binary factor model. To the best of our knowledge, there are only a few literatures on posterior rates of contraction for factor models and its alternative form principal component analysis (PCA). [29] is the first work to consider posterior contraction rates for sparse factor models. [15] derives rate-optimal posterior contraction for sparse PCA. Both results achieve the frequentist minimax rates (up to a logarithmic factor for the first work). Frequentist estimation in factor models include [12], [13] and [14].

Minimax rates for factor models usually appear in the literature in the form of principal component analysis. For example, minimax rates for sparse PCA are derived by [4], [5], [6] and [37] under various settings.

For binary factor models, minimax rates are not available in the literature, and it cannot be easily derived from the existing results. In the current binary factor model setting, there are two main points that deviate from the settings considered in the literature. First, the largest eigenvalue of the matrix $Z_0 Z_0^{\mathrm{T}} + I$ may diverge as $n \to \infty$ in the extreme case, while most minimax rates in the literature for covariance estimation assume bounded spectrum. Second, the binary factor model only takes value in $\{0, 1\}$, which distinguishes itself from ordinary factor models. The results in this paper suggest at least two open problems. First, what is the minimax rate of the binary factor model? Second, is IBP or pIBP rate-optimal? If not, what is the best rate of contraction that can be achieved by the posterior distribution?

7.2 Approximate Group Structure

Theorem 4.4 states the posterior contraction rates of pIBP under the model of a two-group structure through the factor decomposition (5). Such characterization of group structure is exact in the sense that even only one person in S_1 possesses a factor that is mostly possessed by people in S_2, that factor is classified as a common factor, contributing to the total K_0^*. Therefore, in many real cases the exact two-group structure is violated and we can easily get $K_0^* = K_0$, thus losing the advantage of using pIBP.

In this section, we present a result to demonstrate that pIBP still gains advantage over IBP even when $K_0^* = K_0$ but the two-group structure approximately holds. We say Z_0 has an approximate two-group structure if there exists a binary matrix Z^* of the same size such that the number K_0^* associated with Z^* is bounded by $O\left(K_0^{1-\beta}\right)$ and $\|Z_0 Z_0^{\mathrm{T}} - Z^*(Z^*)^{\mathrm{T}}\|_F$ is small. In other words, Z_0 may have a large K_0^*, but it is close to a binary factor matrix whose

K_0^* is small. The following theorem is an oracle inequality for pIBP under the posterior distribution.

Theorem 7.1—*Let* $Z_0 \in \{0,1\}^{n\times K_0}$ *be an arbitrary binary factor matrix, and let* $Z^* \in \{0, 1\}^{n\times K_0}$ be a binary factor matrix with a well specified group structure such that its $K_0^* \lesssim K_0^{1-\beta}$ *for* $\beta \in (0, 1)$. *Under the assumption of Theorem 4.4,*

$$E_{Z_0}\left[\Pi\left(\|ZZ^{\mathrm{T}} - Z_0(Z_0)^{\mathrm{T}}\|_F^2 \le M\left(\frac{n^3K_0^{4-2\beta}}{p} + n^2K_0^2\|Z_0Z_0^T - Z^*(Z^*)^{\mathrm{T}}\|^4\right)|X\right)\right] \ge 1 - exp\left(-C'nK_0^{2(1-\beta)}\right) - \frac{2}{p},$$

for some constants $M, C' > 0$.

In the case when Z_0 has an exact two-group structure, we may choose $Z^* = Z_0$ so that $\|Z_0Z_0^{\mathrm{T}} - Z^*(Z^*)^{\mathrm{T}}\|_F = 0$. Then it reduces to the result in Theorem 4.4. Otherwise, we may choose a Z^* with an exact two-group structure to approximate Z_0. In this case, the posterior distribution contracts to the truth with a rate consisting of two parts. The first part can be viewed as the estimation error of a binary factor matrix Z^* with an exact two-group structure. The second part is the approximation error for the true binary factor matrix Z_0 by Z^*. Note that the rate of convergence for IBP in Theorem 4.3 is $\frac{K_0^4n^3}{p}$. Therefore, as long as

$$n^2K_0^2\|Z_0Z_0^{\mathrm{T}} - Z^*(Z^*)^{\mathrm{T}}\|_F^4 = o\left(\frac{K_0^4n^3}{p}\right),$$

pIBP still converges faster than IBP if the true binary factor matrix Z_0 has an approximate two-group structure.

Let us consider the following example to illustrate Theorem 7.1. Let $Z_0 \in \{0, 1\}^{n\times K_0}$ be a binary factor matrix which generates the data. Among the $K_0 - K_0^* = K_{01} + K_{02}$ factors that possess approximate group structures, there are K_{01} factors belonging to S_1 and K_{02} factors belonging to S_2. In addition, for some small $\delta \in (0, 1)$, n^δ people in S_1 can possess a constant number of factors belonging to S_2, and n^δ people in S_2 can possess a constant number of factors belonging to S_1. We call this situation a δ-approximate two-group structure. By zeroing out these entries, we obtain a binary factor matrix $Z^* \in \{0, 1\}^{n\times K_0}$ with an exact two-group structure, whose factor decomposition is $K_0 = K_{01} + K_{02} + K_0^*$. In other words, for Z^*, there are K_{01} factors exclusively belonging to S_1 and K_{02} factors exclusively belonging to S_2. The approximation error is bounded by $\|Z_0Z_0^{\mathrm{T}} - Z^*(Z^*)^{\mathrm{T}}\|_F^2 \lesssim \|Z_0\|^2\|Z_0 - Z^*\|_F^2 \lesssim n^\delta\|Z_0\|^2$, where $\|\cdot\|$ denotes the spectral norm of a matrix, which is its largest singular value. We summarize this example in the following corollary.

Corollary 7.1—*Under the setting of Theorem 7.1, let* Z^* *have a factor decomposition satisfying* $K_0^* \lesssim K_0^{1-\beta}$, *then as long as* $n^{2\delta} = o\left(\frac{nK_0^2}{p\|Z_0\|^4}\right)$, *we have*

$$E_{Z_0}\left[\Pi\left(\|ZZ^{\mathrm{T}} - Z_0(Z_0)^{\mathrm{T}}\|_F^2 \le \epsilon_{n,p}^2 | X\right)\right] \ge 1 - exp\left(-C' n K_0^{2(1-\beta)}\right) - \frac{2}{p},$$

for some positive sequence $\epsilon_{n,p}^2 = o\left(\frac{K_0^4 n^3}{p}\right)$ *and some constant* $C > 0$.

The corollary provides an example that pIBP converges at a faster rate than that of IBP when Z_0 satisfies the $\mathcal{S}$-approximate two-group structure. The quantity $\|Z_0\|$ quantifies the sparsity of the binary factor matrix Z_0. In many applied situations, the true binary factor matrix Z_0 has a sparse structure [19, 22, 7]. This leads to a small $\|Z_0\|$.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Arriola, Edurne; Marchio, Caterina; Tan, David SP.; Drury, Suzanne C.; Lambros, Maryou B.; Natrajan, Rachael; Maria Rodriguez-Pinilla, Socorro; Mackay, Alan; Tamber, Narinder; Fenwick, Kerry, et al. Genomic analysis of the her2/top2a amplicon in breast cancer and breast cancer cell lines. Laboratory investigation. 2008; 88(5):491–503. [PubMed: 18332872]
2. Barron, Andrew; Schervish, Mark J.; Wasserman, Larry. The consistency of posterior distributions in nonparametric problems. The Annals of Statistics. 1999; 27(2):536–561.
3. Bell D, Berchuck A, Birrer M, Chien J, Cramer DW, Dao F, Dhir R, DiSaia P, Gabra H, Glenn P, et al. Integrated genomic analyses of ovarian carcinoma. Nature. 2011; 474:609–615. [PubMed: 21720365]
4. Birnbaum, Aharon; Johnstone, Iain M.; Nadler, Boaz; Paul, Debashis. Minimax bounds for sparse pca with noisy high-dimensional data. Annals of statistics. 2013; 41(3):1055. [PubMed: 25324581]
5. Tony Cai T, Ma Zongming, Wu Yihong. Sparse pca: Optimal rates and adaptive estimation. The Annals of Statistics. 2013; 41(6):3074–3110.
6. Cai, Tony; Ma, Zongming; Wu, Yihong. Optimal estimation and rank detection for sparse spiked covariance matrices. Probability Theory and Related Fields. 2013:1–35.
7. Carvalho, Carlos M.; Chang, Jeffrey; Lucas, Joseph E.; Nevins, Joseph R.; Wang, Quanli; West, Mike. High-dimensional sparse factor modeling: applications in gene expression genomics. Journal of the American Statistical Association. 2008; 103(484):1438–1456. [PubMed: 21218139]
8. Cerami, Ethan; Gao, Jianjiong; Dogrusoz, Ugur; Gross, Benjamin E.; Onur Sumer, Selcuk; Arman Aksoy, Bülent; Jacobsen, Anders; Byrne, Caitlin J.; Heuer, Michael L.; Larsson, Erik, et al. The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer discovery. 2012; 2(5):401–404. [PubMed: 22588877]
9. Chen, Mengjie; Gao, Chao; Zhao, Hongyu. Posterior contraction rates of the phylo-genetic indian buffet processes. 2014
10. Chène, Patrick. Inhibiting the p53–mdm2 interaction: an important target for cancer therapy. Nature reviews cancer. 2003; 3(2):102–109. [PubMed: 12563309]
11. Diaconis, Persi; Freedman, David. On the consistency of bayes estimates. The Annals of Statistics. 1986:1–26.
12. Fan, Jianqing; Fan, Yingying; Lv, Jinchi. High dimensional covariance matrix estimation using a factor model. Journal of Econometrics. 2008; 147(1):186–197.
13. Fan, Jianqing; Liao, Yuan; Mincheva, Martina. High dimensional covariance matrix estimation in approximate factor models. Annals of statistics. 2011; 39(6):3320. [PubMed: 22661790]

14. Fan, Jianqing; Liao, Yuan; Mincheva, Martina. Large covariance estimation by thresholding principal orthogonal complements. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2013; 75(4):603–680.
15. Gao, Chao; Zhou, Harrison H. Rate-optimal posterior contraction for sparse pca. Annals of Statistics, to appear. 2015
16. Ghosal, Subhashis; Van Der Vaart, Aad. Convergence rates of posterior distributions for noniid observations. The Annals of Statistics. 2007; 35(1):192–223.
17. Ghosal, Subhashis; Ghosh, Jayanta K.; van der Vaart, Aad W. Convergence rates of posterior distributions. Annals of Statistics. 2000; 28(2):500–531.
18. Griffiths, Thomas L.; Ghahramani, Zoubin. In NIPS. MIT Press; 2005. Infinite latent feature models and the indian bu et process.; p. 475-482.
19. Griffiths, Thomas L.; Ghahramani, Zoubin. The indian buffet process: An introduction and review. Journal of Machine Learning Research. 2011; 12:1185–1224.
20. Hu, Guohong; Chong, Robert A.; Yang, Qifeng; Wei, Yong; Blanco, Mario A.; Li, Feng; Reiss, Michael; Au, Jessie L-S.; Haffty, Bruce G.; Kang, Yibin. Mtdh activation by 8q22 genomic gain promotes chemoresistance and metastasis of poor-prognosis breast cancer. Cancer cell. 2009; 15(1):9–20. [PubMed: 19111877]
21. Killian, Audrey; Sarafan-Vasseur, Nasrin; Sesboüé, Richard; Le Pessot, Florence; Blanchard, France; Lamy, Aude; Laurent, Michelle; Flaman, Jean-Michel; Frébourg, Thierry. Contribution of the bop1 gene, located on 8q24, to colorectal tumorigenesis. Genes, Chromosomes and Cancer. 2006; 45(9):874–881. [PubMed: 16804918]
22. Knowles, David; Ghahramani, Zoubin. Nonparametric bayesian sparse factor models with application to gene expression modeling. The Annals of Applied Statistics. 2011; 5(2B):1534–1552.
23. Le Cam, Lucien; Lo Yang, Grace. Asymptotics in statistics: some basic concepts. Springer; 2000.
24. Lv, Qi; Wang, Wei; Xue, Jianfei; Hua, Fang; Mu, Rong; Lin, Heng; Yan, Jun; Lv, Xiaoxi; Chen, Xiaoguang; Hu, Zhuo-Wei. Dedd interacts with pi3kc3 to activate autophagy and attenuate epithelial–mesenchymal transition in human breast cancer. Cancer research. 2012; 72(13):3238–3250. [PubMed: 22719072]
25. Miller, Jeffrey W.; Harrison, Matthew T. Inconsistency of pitman-yor process mixtures for the number of components. arXiv preprint arXiv. 2013; 1309.0024
26. Miller, Kurt T.; Griffiths, Thomas; Jordan, Michael I. The phylogenetic indian buffet process: A non-exchangeable nonparametric prior for latent features. arXiv preprint arXiv. 2012; 1206.3279
27. Muzny, Donna M.; Bainbridge, Matthew N.; Chang, Kyle; Dinh, Huyen H.; Drummond, Jennifer A.; Fowler, Gerald; Kovar, Christie L.; Lewis, Lora R.; Morgan, Margaret B.; Newsham, Irene F., et al. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012; 487:330–337. [PubMed: 22810696]
28. Nik-Zainal, Serena; Van Loo, Peter; Wedge, David C.; Alexandrov, Ludmil B.; Greenman, Christopher D.; Wai Lau, King; Raine, Keiran; Jones, David; Marshall, John; Ramakrishna, Manasa, et al. The life history of 21 breast cancers. Cell. 2012; 149(5):994–1997. [PubMed: 22608083]
29. Pati, Debdeep; Bhattacharya, Anirban; Pillai, Natesh S.; Dunson, David. Posterior contraction in sparse bayesian factor models for massive covariance matrices. The Annals of Statistics. 2014; 42(3):1102–1130.
30. Ramakrishna, Manasa; Williams, Louise H.; Boyle, Samantha E.; Bearfoot, Jennifer L.; Sridhar, Anita; Speed, Terence P.; Gorringe, Kylie L.; Campbell, Ian G. Identification of candidate growth promoting genes in ovarian cancer through integrated copy number and expression analysis. PloS one. 2010; 5(4):e9983. [PubMed: 20386695]
31. Rousseau, Joel; Têtu, Bernard; Caron, Danielle; Malenfant, Patrick; Cattaruzzi, Paola; Audette, Marie; Doillon, Charles; Tremblay, Jacques P.; Guérette, Benoit. Rcas1 is associated with ductal breast cancer progression. Biochemical and biophysical research communications. 2002; 293(5): 1544–1549. [PubMed: 12054692]
32. Schwartz, Lorraine. On bayes procedures. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete. 1965; 4(1):10–26.

33. TCGA. Comprehensive molecular portraits of human breast tumours. Nature. 2012; 490:61–70. [PubMed: 23000897]

34. Whye Teh, Yee; Görür, Dilan; Ghahramani, Zoubin. Stick-breaking construction for the indian buffet process. Proceedings of the International Conference on Artificial Intelligence and Statistics. 2007; 11

35. Van De Vijver, Marc J.; He, Yudong D.; van't Veer, Laura J.; Dai, Hongyue; Hart, Augustinus AM.; Voskuil, Dorien W.; Schreiber, George J.; Peterse, Johannes L.; Roberts, Chris; Marton, Matthew J., et al. A gene-expression signature as a predictor of survival in breast cancer. New England Journal of Medicine. 2002; 347(25):1999–2009. [PubMed: 12490681]

36. Vershynin, Roman. Introduction to the non-asymptotic analysis of random matrices. arXiv preprint arXiv. 2010; 1011.3027

37. Vu, Vincent Q.; Lei, Jing. Minimax sparse principal subspace estimation in high dimensions. The Annals of Statistics. 2013; 41(6):2905–2947.

38. Wrzeszczynski, Kazimierz O.; Varadan, Vinay; Byrnes, James; Lum, Elena; Kamalakaran, Sitharthan; Levine, Douglas A.; Dimitrova, Nevenka; Zhang, Michael Q.; Lucito, Robert. Identification of tumor suppressors and oncogenes from genomic and epigenetic features in ovarian cancer. PLoS One. 2011; 6(12):e28503. [PubMed: 22174824]

39. Zhang, Dingxiao; Jiang, Pingping; Xu, Qinqin; Zhang, Xiaoting; Zhang, Dingxiao; Jiang, Pingping; Xu, Qinqin; Zhang, Xiaoting. 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. 2011; 286(20):17746–17754. [PubMed: 21454576]

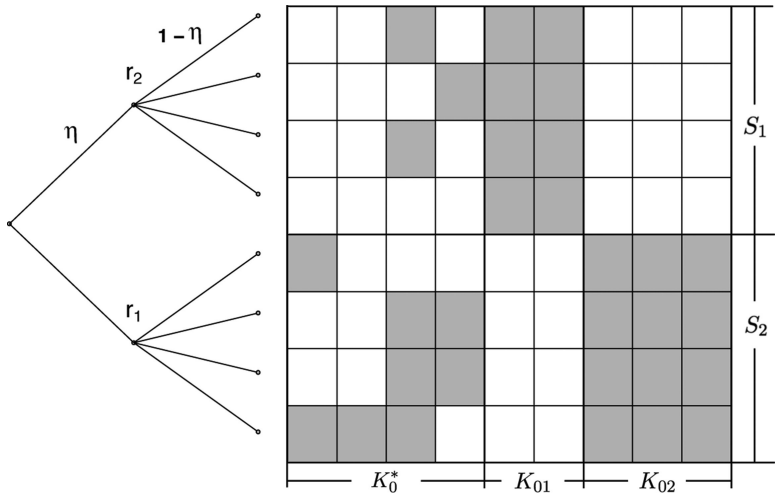

Figure 1.
An illustration of the two group tree and the factor decomposition.

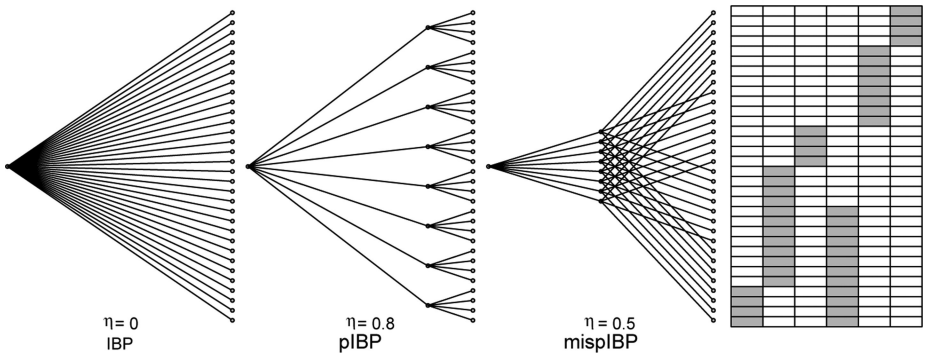

Figure 2.

The illustration of 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.

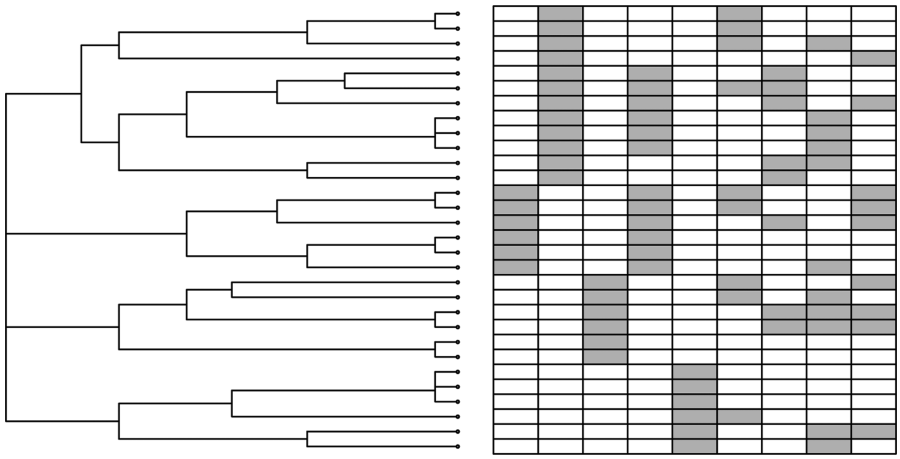

Figure 3.
The illustration of the latent factor matrix Z_0 and tree prior constructed from the hierarchical clustering analysis of Z_0 in the second simulation.

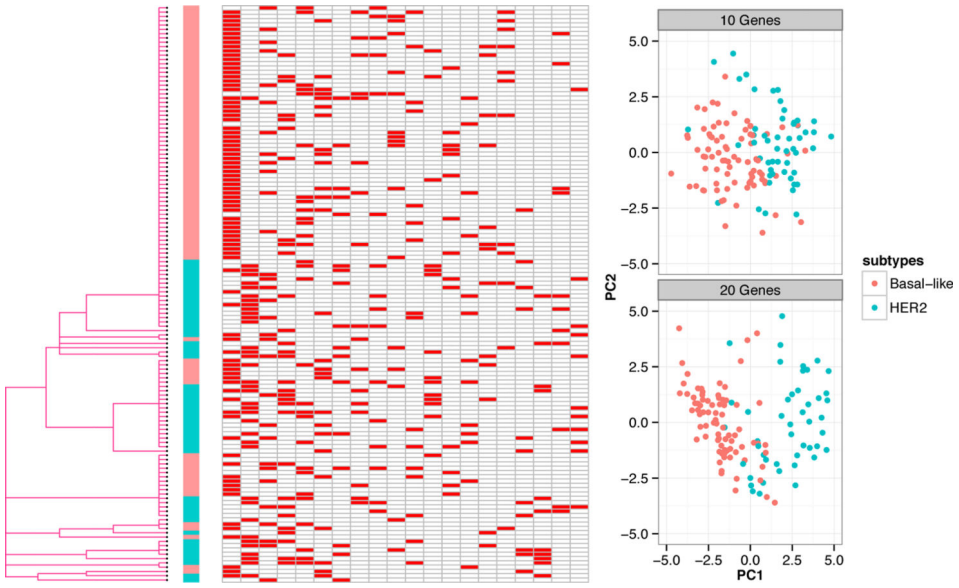

Figure 4.

A graph showing the 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.

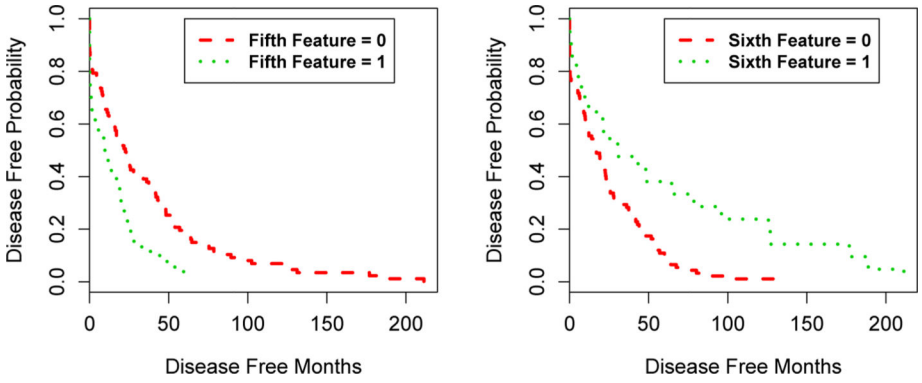

Figure 5.

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

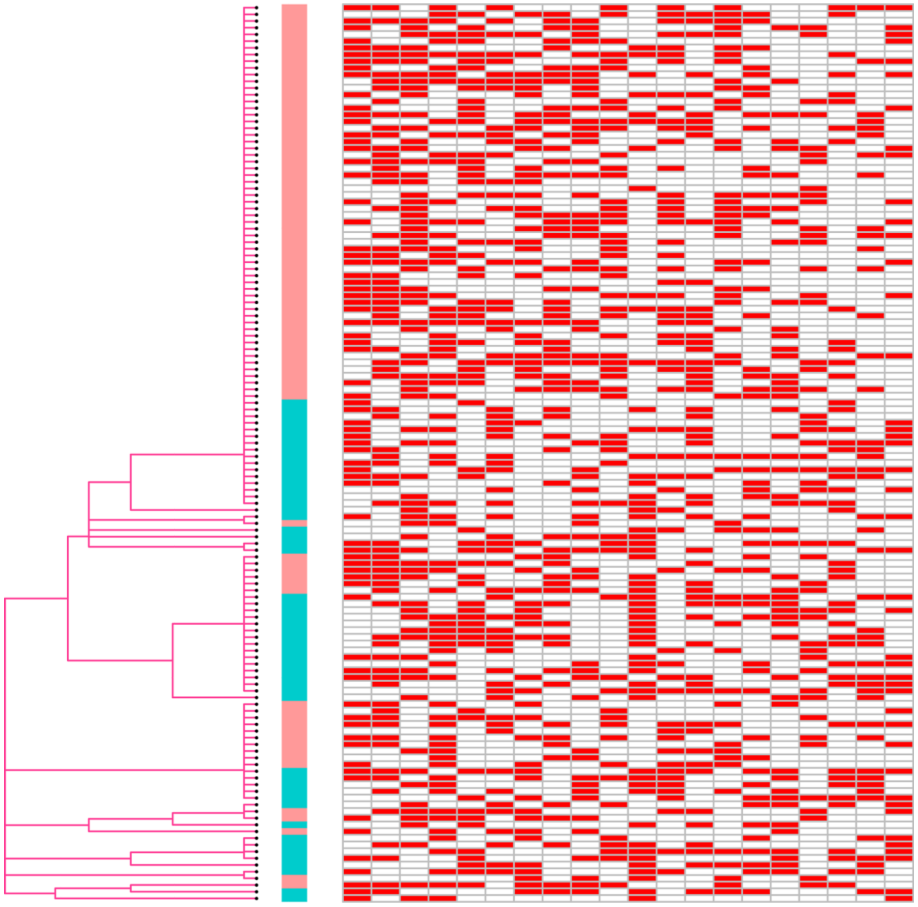

Figure 6.

IBP result on TCGA breast cancer samples. This plot shows the dendrogram tree prior (left), the inferred latent factor matrix Z (right, only first 20 columns shown) and subtype status (middle, Basal-like as Red and HER2 as Green).

Table 1

Simulation results: comparisons of IBP, pIBP with the appropriate tree prior and pIBP with the mis-specified tree prior (mispIBP).

(*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)

The performance is measured by estimation errors in terms of the normalized Frobenius norm of the feature similarity matrix $n^{-1/2}\| ZZ^{\mathrm{T}} - Z_0 Z_0^{\mathrm{T}} \|_F$ (F-norm), and the number of estimated latent factors $\hat{K}$. Numbers in parentheses are the standard deviations across the 40 independent replicates. In the above models, $\sigma^2_{A,0}=1, \sigma^2_{X,0}=0 \cdot 5$, $K_0 = 9$, results are based on 1000 Markov chain Monte Carlo steps.

Table 2

Simulation results: comparisons of IBP and pIBP with the tree prior from the dendrogram of a hierarchical clustering on Z_0.

(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)

The performance is based on 40 independent replicates, each with 1000 Markov chain Monte Carlo steps.