Published in final edited form as:

Stat Appl Genet Mol Biol. 2017 April 25; 16(2): 145–158. doi:10.1515/sagmb-2016-0051.

A Bayesian semiparametric factor analysis model for subtype identification

Jiehuan Sun,

Department of Biostatistics, Yale School of Public Health, 60 College Street, New Haven, CT 06510, USA

Joshua L. Warren, and

Department of Biostatistics, Yale School of Public Health, 60 College Street, New Haven, CT 06510, USA

Hongyu Zhao*

Department of Biostatistics, Yale School of Public Health, 60 College Street, New Haven, CT 06510, USA

Abstract

Disease subtype identification (clustering) is an important problem in biomedical research. Gene expression profiles are commonly utilized to infer disease subtypes, which often lead to biologically meaningful insights into disease. Despite many successes, existing clustering methods may not perform well when genes are highly correlated and many uninformative genes are included for clustering due to the high dimensionality. In this article, we introduce a novel subtype identification method in the Bayesian setting based on gene expression profiles. This method, called BCSUB, adopts an innovative semiparametric Bayesian factor analysis model to reduce the dimension of the data to a few factor scores for clustering. Specifically, the factor scores are assumed to follow the Dirichlet process mixture model in order to induce clustering. Through extensive simulation studies, we show that BCSUB has improved performance over commonly used clustering methods. When applied to two gene expression datasets, our model is able to identify subtypes that are clinically more relevant than those identified from the existing methods.

Keywords

Bayesian factor analysis; Bayesian nonparametrics; clustering; Dirichlet process; gene expression study

1 Introduction

For complex diseases such as cancers, there exist great heterogeneities among patients (e.g. Jester and Geschwind, 2014), which are often attributed to disease subtypes. Many studies have been conducted to identify disease subtypes in the past decades, largely due to the

*Corresponding author: Hongyu Zhao, Department of Biostatistics, Yale School of Public Health, 60 College Street, New Haven, CT 06510, USA, hongyu.zhao@yale.edu.

advancement of molecular technologies, such as gene expression microarrays (Sorlie et al., 2001; Wigle et al., 2002). Disease subtypes identified using gene expression data are usually associated with clinical features and/or outcomes and hence they are valuable in understanding disease onset and progression (Sorlie et al., 2001; Wigle et al., 2002).

Existing approaches of inferring disease subtypes based on gene expression profiles can be broadly divided into two categories, where they share the common first step in selecting important genes related to the disease under study. After gene selections, standard clustering methods are used directly on those selected genes to infer subtypes for approaches in the first category, such as K-means (Hartigan and Wong, 1979), partitioning around medoids method (PAM) (Cali ski and Harabasz, 1974), model-based clustering (MCLUST) (Fraley and Raftery, 2002), and hierarchical clustering (Johnson, 1967). However, these approaches might not perform well when the variables used for clustering are highly correlated and contain a large proportion of uninformative genes, as commonly observed in gene expression data. To deal with these challenges, approaches in the second category first use matrix factorization methods, such as principal component analysis (PCA) (Wold et al., 1987) and non-negative matrix factorization (NMF) (Lee and Seung, 2001), to define metagenes and then clustering methods are applied on the metagenes to infer subtypes. It has been shown that the metagene-based approaches can usually improve clustering results by reducing noise and allowing for improved visualizations and biologically meaningful interpretations (Tamayo et al., 2007; Liu et al., 2008). However, all current metagene-based approaches are two-step procedures, where the definition of metagenes and subsequent clustering are conducted in two separate steps. As a result, the uncertainties associated with estimating the metagenes are ignored during clustering.

A factor analysis model, where a few unobserved latent factors are used to model multivariate data, is commonly used for dimension reduction in genomics studies (West, 2003; Carvalho et al., 2008). Similar to PCA and NMF, the factor analysis model is also a matrix factorization method, which tries to preserve the correlation structure in the data. In the factor analysis model, the factor loading matrix can be used to define metagenes while the factor scores can be seen as the expression values of the metagenes and used for clustering.

A Dirichlet process mixture model (DPM) is an infinite mixture model that has been previously used for clustering gene expression data (e.g. Dahl, 2006; Kim et al., 2006; Qin, 2006). Compared to finite mixture models, such as MCLUST, DPM does not require specification of the number of mixture components a priori and the clustering structure obtained from DPM is in fact averaging over models of all possible numbers of mixture components and hence is robust to the misspecification of the number of clusters. Moreover, DPM also allows for the assessment of the uncertainty in the number of clusters. However, sampling from DPM is challenging for high dimensional data, where specific techniques, such as the split-merge algorithm, might be needed (Jain and Neal, 2004).

In this article, we propose a metagene-based clustering method for disease subtype identification under the Bayesian framework, called Bayesian Clustering method for Subtype identification (BCSub), which defines metagenes and identifies subtypes

simultaneously in a single step. More specifically, BC-Sub adopts the factor analysis model to construct metagenes and then the Dirichlet process (DP) prior (Ferguson, 1973; Sethuraman, 1994) is specified on factor scores to induce clustering. Compared to the existing two-step methods, BCSub can appropriately account for the uncertainties in defining metagenes when performing clustering and hence improve clustering performance over two-step procedures. In addition, the adoption of the factor analysis model in BCSub reduces the dimension of the data to a few factor scores, which not only alleviates the impact of high correlations among genes and the un-informative genes included for clustering, but also allows efficient sampling from the DPM. An R package to implement the newly developed method is available from CRAN with package name “BCSub” (<https://cran.r-project.org/web/packages/BCSub/index.html>).

The rest of this article is organized as follows. Section 2 presents the details of our proposed method. Section 3 displays prior specifications and computation details. Section 4 shows the comparison of clustering performance for different methods in simulation studies. Section 5 presents the comparisons of all methods in real data applications. We conclude the paper in Section 6.

2 Methods

2.1 Statistical model

In this section, we describe our proposed model and inferential procedure. Let $[Y_i]_{G \times 1}$ denote the gene expression profile vector for the i^{th} subject, where $i = 1, 2, \dots, n$, n is the number of subjects, and G is the total number of genes. Then, a factor analysis model with M factors is introduced to model the observed expression data such that

$$Y_i = \Lambda \eta_i + \varepsilon_i, \quad \varepsilon_i \stackrel{\text{iid}}{\sim} \text{MVN}(\mathbf{0}_G, \Sigma) \quad (1)$$

where $[\eta_i]_{M \times 1}$ is the vector of factor scores for subject i , $[\Lambda]_{G \times M}$ is the factor loading matrix, $\mathbf{0}_G$ is a vector of length G with all elements being 0, and $[\Sigma]_{G \times G}$ is a diagonal covariance matrix with diagonal elements being σ_g^2 , $g = 1, 2, \dots, G$. Here, we assume that the data matrix is standardized across all samples and hence the mean is assumed to be 0 for Y_i in Equation (1).

In order to ensure identifiability, constraints have to be placed on the factor loading matrix Λ (Geweke and Zhou, 1996; Erosheva and Curtis, 2011). Different constraints can be imposed on Λ , which do not influence the model fit (Erosheva and Curtis, 2011). We constrain Λ to be a lower-triangular matrix with all diagonal elements being one (Erosheva and Curtis, 2011) so that we can model the covariance matrix of the factor scores flexibly.

Then, the DPM is assumed for the factor scores η_i 's such that

$$\begin{aligned}\eta_i|F, \Omega &\sim \int \Phi(\eta_i; \boldsymbol{\mu}, \Omega) F(d\boldsymbol{\mu}), \\ F &\sim DP(c, F_0),\end{aligned}\quad (2)$$

where F is the unknown distribution for the means of the factor scores, $\Phi(\cdot; \boldsymbol{\mu}, \Omega)$ is the density function for a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and diagonal covariance matrix Ω , c and F_0 are the concentration parameter and base distribution, respectively, in the DP prior. In our case, the base distribution F_0 is taken to be $MVN(\mathbf{0}_M, \rho \mathbf{I}_M)$, where $\mathbf{I}_M$ is the identity matrix with dimension M and ρ controls the overall variability in all factor scores.

The model is fully specified after prior distributions are selected for all unknown model parameters (see Section 3.1 for details). When appropriate, we choose conjugate priors for most parameters and hence it is straightforward to apply Gibbs sampling to obtain posterior samples through a Markov chain Monte Carlo (MCMC) algorithm. The infinite number of mixture components in the DPM models makes the posterior sampling computationally difficult in practice. Many sampling algorithms have been proposed to address this challenge, including Neal (1992); MacEachern (1994); Walker (2007); Papaspiliopoulos and Roberts (2008) and Kalli et al. (2011). In our work, we adopt the efficient method proposed by Neal (1992) and MacEachern (1994) for sampling, where the distribution of the parameters involved in the mixture components is integrated out over its prior distribution and, as a result, those parameters follow a generalized Pólya urn scheme.

2.2 Choosing the number of factors

The number of factors in the factor analysis model is unknown a priori and is challenging to infer in practice (Hoyle and Duvall, 2004). In the standard Bayesian factor analysis model, a few methods have been proposed to determine the number of factors (see Lopes and West (2004) and references therein). Among these, the method proposed in Lopes and West (2004) is a popular one, where the number of factors is treated as an unknown parameter and reversible jump MCMC algorithm is used to update this parameter. However, the factor analysis model is coupled with the DPM model in our method and the performance of these proposed techniques in our framework is unknown. Moreover, incorporation of the inference on the number of factors in the method can increase computation time considerably. To improve computational efficiency, especially in cases where a large number of factors might be needed (e.g. when the number of genes (G) is large and the correlation structure among genes is complex), we propose to first select the number of factors and then apply our proposed model for the given numbers of factors. Specifically, we use the *parallel* function (Drasgow and Lissak, 1983) within the R package “nFactors” to infer the number of factors, which is based on the empirical distribution of the eigenvalues of the correlation matrix of uncorrelated normal variables.

2.3 Clustering inference

In Bayesian mixture models, the labels for different clusters can exchange during the MCMC sampling without changing the likelihood and hence it is meaningless to summarize

the clustering results by using the ergodic averages of the labels (Jasra et al., 2005). Instead, the posterior similarity matrix, which can be easily calculated based on the collected MCMC samples, is usually used to derive the clustering structure, since it is robust to the label switching issues. The $(i, j)^{\text{th}}$ entry of the posterior similarity matrix, which represents the posterior probability that subject i and subject j are in the same cluster, can be calculated

from the S posterior samples as $\frac{1}{S} \sum_{s=1}^S \delta(e_i^{(s)} = e_j^{(s)})$, where $\delta(\cdot)$ is the indicator function and $e_i^{(s)}$ and $e_j^{(s)}$ represent the cluster membership for the i^{th} and j^{th} subjects during the s^{th} iteration.

Several methods have been proposed to infer the optimal number of clusters and the clustering structure based on the posterior similarity matrix, including Binder (Binder, 1978), Dahl's criterion (Dahl, 2006), and MPEAR (Fritsch et al., 2009). Among these, the objective function that MPEAR optimizes over is most appealing. It optimizes over the expectation of the adjusted Rand index between the estimated and true clustering structure. The adjusted Rand index represents the degree of agreement between two clustering partitions (it usually ranges from 0 to 1 with larger value indicating better agreement, but can also be negative in some situations) and is a commonly used measure of clustering performance (Hubert and Arabie, 1985; Yeung and Ruzzo, 2001; Lock and Dunson, 2013). Therefore, we adopt MPEAR with average linkage in our simulation studies.

Another approach, as introduced in Medvedovic et al. (2004), to inferring the clusters is the classical agglomerative hierarchical clustering method (HCLUST). In this approach, the posterior similarity matrix is used to generate the pairwise distance of all subjects and then the pairwise distance is given as input to HCLUST with average linkage to infer the clustering structure for a given number of clusters. This approach is useful when the number of clusters is known a priori or can be determined based on clinical relevance for better interpretation. Also, since the number of clusters could affect the adjusted Rand index, which is usually unknown in real data, we compare different clustering methods by fixing the number of clusters and hence we use this approach in the real data analysis.

2.4 Methods for comparison

To demonstrate the usefulness of our proposed method, we compare the clustering performance of BCSUB with four commonly used clustering methods, including PAM, MCLUST, and two two-step procedures based on NMF and PCA, respectively.

The PAM method (Calinski and Harabasz, 1974) is a robust alternative to K-means, since it minimizes a sum of dissimilarities instead of a sum of squared euclidean distances. The optimal number of clusters is determined by average silhouette width in PAM (Rousseeuw, 1987). We use the *pamk* function in R package “fpc” in our comparisons.

The MCLUST method (Fraley and Raftery, 2002), a finite Gaussian mixture model, optimizes over different numbers of clusters and different structures for the covariance matrix using the expectation-maximization algorithm. The optimal number of clusters and covariance matrix structure is determined based on the Bayesian information criterion. We use the *Mclust* function in R package “mclust” in our analysis.

For the two-step procedure based on PCA, PCA is first applied to obtain principal component scores for each subject and then MCLUST is used for clustering on the principal component scores. The number of principal components retained is determined using the parallel analysis as in our newly introduced method. PCA is implemented using the function *prcomp* in R. We will denote this method as PCA in our following discussion.

The two-step procedure based on NMF is implemented similarly as that based on PCA. Since NMF requires all elements in the data matrix to be nonnegative, all data matrices are transformed to nonnegative matrices whenever this approach is used. The NMF is implemented using the *nmf* function in the R package “NMF”. We will denote this method as NMF in our following discussion.

3 Inference

In this section, we describe our prior specifications for unknown model parameters and posterior sampling details in both simulation studies and real data applications.

3.1 Prior specification

The concentration parameter c in the DP prior controls the prior expectation of the number of clusters in the data, where a larger value of c leads to a larger prior expected number of clusters. We fix $c = 1$ for our analyses, which results in the prior expected number of clusters to be 4.49 for 50 subjects based on Theorem 1 in Rodriguez and Dunson (2014). Therefore, $c = 1$ provides a reasonable prior expected number of clusters, since we usually do not expect many subtypes given a disease under study.

We specify independent and weakly informative conjugate prior distributions for all variance parameters involved in the diagonal matrices Σ and Ω through the Inverse Gamma(0.01, 0.01) distribution. We select a Uniform(0, 2) prior distribution for the variance parameter ρ in the base distribution of the DP prior, where the upper bound is reasonably large, since the data matrix is standardized prior to clustering analysis.

Next, we specify prior distributions for the unknown parameters in the factor analysis model. Independent normal prior distributions $N(0, \sigma^2)$ are used for all off-diagonal elements of the loading matrix Λ . In our application, we choose $\sigma^2 = 1$, which is reasonable for standardized data matrices.

All priors other than ρ are selected to be conjugate. Hence, the Gibbs sampling algorithm can be used to sample from the posterior distributions of all parameters other than ρ . For ρ , we use a Metropolis Hastings algorithm for posterior sampling. See Appendix for more details on the sampling algorithm.

3.2 Computational details

In simulation studies, 20,000 MCMC samples are generated and the first 10,000 samples are discarded as burn-in. All subjects are initially placed in one cluster. The initial value for ρ is set to be 1. The initial value for Σ is taken to be identity matrix I_G . The initial values for Λ and η_i 's are randomly sampled from their prior distributions.

For the real data analyses, 30,000 samples are collected after a burn-in period of 30,000 iterations from three separate chains for a total of 90,000 posterior samples available for making posterior inference. Starting values for each chain are selected to be over-dispersed with respect to the target marginal posterior distributions based on early runs of the model. Convergence monitoring is conducted by visual inspection of all traceplots and by calculating the differences among the posterior similarity matrices given by the three chains, since our primary goal is clustering as determined by the posterior similarity matrix.

4 Simulation studies

In this section, the performance of BCSub, PAM, MCLUST, PCA, and NMF is evaluated in simulation studies. The true cluster membership is known for all simulation settings and hence the adjusted Rand index is calculated directly to indicate clustering performance.

4.1 Simulation setup

In the simulation studies, we assess the influence of both uninformative genes and different covariance structures on the performance of these clustering methods. In all simulation settings, we generate expression profiles of 200 genes for 100 subjects coming from two clusters, i.e. 50 subjects per cluster. Here, we assume that the genes are pre-filtered prior to cluster analysis to select informative ones. Also, we try to make the clustering task challenging so as to distinguish the performance of different clustering methods. If many genes are informative of clustering, any clustering method could easily identify the true clustering structure well. Hence, only 200 genes are considered in our simulations. The gene expression profiles for the subjects in two clusters are generated from multivariate normal distributions with different means and the same covariance matrix, that is

$$Y_i|e_i=1 \sim \text{MVN}(\mu_1, \Sigma), \quad (3)$$

$$Y_i|e_i=2 \sim \text{MVN}(\mu_2, \Sigma), \quad (4)$$

where different configurations of μ_1 , μ_2 , and Σ are employed as detailed below in order to assess the influence of both uninformative genes and different covariance structures on the clustering performance.

To assess the influence of the covariance structures, we consider two different covariance matrices, which define two simulation settings “F” and “R” as follows.

- F: we construct the covariance matrix based on a standard factor analysis model, that is $\Sigma = \Lambda\Lambda^T + I_{200}$, where Λ is a factor loading matrix with four factors. Let

the vector λ be
$$\lambda = [\underbrace{\lambda_1, \dots, \lambda_{50}}_{\text{Block 1}}, \underbrace{\lambda_{51}, \dots, \lambda_{100}}_{\text{Block 2}}, \underbrace{\lambda_{101}, \dots, \lambda_{150}}_{\text{Block 3}}, \underbrace{\lambda_{151}, \dots, \lambda_{200}}_{\text{Block 4}}]^T.$$
 Then, the four factor loadings in Λ are taken to be λ with the first factor loading

only having non-zero values in block 1, the second factor loading only having non-zero values in block 2, and so on, where all non-zero values are generated from the $\text{Uniform}(-3, 3)$ distribution.

- R: to more closely represent real data, we consider another covariance matrix $\Sigma = \hat{R}$, which is the estimated correlation matrix based on the top 200 genes with the largest variances in the breast cancer data (see Section 5.1 for a detailed description).

For simulation setting F, we consider $\mu_1 = \underbrace{[0.5, \dots, 0.5]}_{200}^T$. To study the influence of uninformative genes, we consider two other scenarios, where 50 elements and 100 elements in μ_1 are randomly selected and set to be 0, respectively. μ_2 is fixed to be a vector of zeros for all three scenarios. We denote the three scenarios as {F0, F50, F100}. Similarly, we have {R0, R50, R100} for simulation setting R, except that the non-zero elements in μ_1 are changed to 1.0. The simulation setting F allows us to assess the performance of the clustering methods as well as the estimation of the number of factors while setting R allows us to assess the performance of the clustering methods in a situation closely resembling our specific dataset.

4.2 Simulation results

For each scenario, we simulate 50 datasets and the adjusted Rand index is calculated for each clustering method on each of the simulated datasets. The number of factors is estimated using the parallel analysis for each dataset.

The means of the estimated number of factors across all 50 datasets using the parallel analysis are {5.0, 5.0, 5.0} and {7.1, 8.4, 9.0} for settings {F0, F50, F100} and {R0, R50, R100}, respectively. It seems that the parallel analysis tends to overestimate the number of factors in setting F, where the true number of factors is 4. This is potentially due to the fact that the parallel analysis ignores the clustering structure in the data. In addition, the estimated numbers of factors are different for different scenarios under simulation setting R, which suggests that the proportion of uninformative genes also affects the estimation of the number of factors.

The averages and standard deviations of the adjusted Rand index over 50 datasets are shown in Table 1, from which we can see that the performance of BCSUB is consistently better than or similar to other methods in all scenarios. Comparing the clustering methods in the second category, including BCSUB, PCA, and NMF, we find that BCSUB has better or similar performance than the other two methods in all scenarios. PCA performs reasonably well in setting F, since the underlying covariance matrix is low rank and matches the assumption of PCA. However, the performance of PCA is extremely poor in setting R, which may be caused by the estimated correlation matrix from the real data that is not low rank. In contrast, it seems that the performance of BCSUB is robust in this setting. Overall, NMF does not perform well, which may be due to the non-negativity constraints on the data matrix (we have to arbitrarily transform the standardized data matrix to a non-negative matrix) and also

the method of determining the number of factors (the parallel analysis is mainly developed for PCA and factor analysis models).

Next, we compare the performance of BCSUB to the clustering methods in the first category, i.e. MCLUST and PAM. When the proportion of the uninformative genes increases, the performance of MCLUST and PAM becomes worse. However, it seems BCSUB is robust to the uninformative genes, especially in setting F, where the covariance matrix is low rank. This suggests that the dimension reduction property embedded in BCSUB could reduce noise and improve clustering. Note that the difference in the means of gene expression profiles for the two clusters in setting R is larger than that in setting F and hence all other methods have improved performance in setting R. However, this is not true for BCSUB, suggesting that BCSUB performs best when the underlying covariance matrix is truly low rank. However, BCSUB also outperforms the other methods in setting R.

Based on the simulation results, BCSUB has improved performance over competing clustering methods. This is especially true in the situation where the selected genes are highly correlated, which can be explained by a few latent factors, and some of the selected genes are uninformative of the clustering; a commonly observed setting in gene expression data.

5 Real data analysis

In this section, we apply BCSUB, PCA, PAM, and MCLUST to two real datasets and compare their performance. The NMF method is omitted for comparison due to its poor performance in simulations. The true cluster membership is unknown in the real data and hence important clinical variables are used to define clinically relevant subtypes, based on which the adjusted Rand index is calculated. In each of our analyses, the data matrix is standardized prior to clustering analysis as commonly done in the literature (Pan and Shen, 2007). Here, we consider two datasets, where sample size is large and the clustering task is relatively difficult so that it is easier to see the differences in the clustering performance of different methods. All methods should work well in discriminating cancer patients with normal controls and hence we don't consider such datasets.

5.1 Breast cancer data

We first compare the clustering methods on the Agilent mRNA expression microarrays data of breast cancer patients from The Cancer Genome Atlas (TCGA) project (The Cancer Genome Atlas Network, 2012). It is well accepted that several subtypes exist in breast cancer patients, where PAM50 gene signatures are commonly used to define the subtypes, which are highly associated with several biomarkers, including oestrogen receptor (ER), progesterone receptor (PR), and HER2 (Perou et al., 2000; Parker et al., 2009; The Cancer Genome Atlas Network, 2012). The normalized gene expression data of the 531 breast cancer patients are downloaded from the UCSC cancer browser (<https://genome-cancer.ucsc.edu>) and the 622 genes with the largest variances (standard deviation greater than 1.5) are selected for clustering. The number of factors given by the parallel analysis on this dataset is 30.

First, in order to see how well these clustering methods could identify the clinically relevant subtypes defined by the PAM50, we treat the subtypes defined in The Cancer Genome Atlas Network (2012) as true clustering structure and calculate the adjusted Rand index. The subtypes include 96 Basal-like, 58 HER2-enriched, 231 Luminal A, 127 Luminal B, and 8 Normal-like (11 unclassified patients are omitted). Since the number of clusters could affect the adjusted Rand index and the number of the true underlying subtypes is unknown, we choose to compare these methods by fixing the number of clusters. Specifically, we calculate the adjusted Rand index for each method given a pre-specified number of clusters that ranges from 2 to 10 (Here, HCLUST is used to derive the clustering structure for BCSUB). From Figure 1A, we can see that BCSUB outperforms the other methods for most of the given numbers of clusters including the number of clusters being 5 (The number of PAM50 defined subtypes is 5).

Next, we want to see if the subtypes defined by different methods are associated with ER, PR, and HER2, since these are important biomarkers for breast cancer and higher association with these biomarkers indicate greater clinical relevance of the defined subtypes. Based on the ER status, the patients can be divided into three different subtypes with 404 ER positive, 118 ER negative, and two indeterminate (Seven patients with no ER or PR status are omitted). As shown in Figure 1B, we can see that BCSUB has much better performance than the other methods, except when the number of clusters is 2, in which case MCLUST performs the best. We observe similar results for the subtypes defined based on PR status (342 PR positive, 179 PR negative, and three indeterminate), as shown in Figure 2A. The performance of the other methods decreases dramatically when the specified number of clusters changes from 2 to 3, which may be due to the small number of patients in the indeterminate group and that the other methods tend to separate the patients into clusters of similar sizes. Thus, the performance of the other methods could be impaired by the misspecification of the number of clusters and the imbalanced number of patients across subtypes while it seems that BCSUB is robust to the misspecification in a certain range of number of clusters. For HER2 status, all methods fail to achieve reasonable concordance with the subtypes defined by HER2 (Result is not shown here). One reason for this could be that HER2 is not specific to any subtypes defined by PAM50, as observed in The Cancer Genome Atlas Network (2012), and hence might not be a perfect biomarker alone to define subtypes that have distinctive gene expression profiles.

Moreover, BCSUB identifies one interesting cluster that is totally missed by other methods or in the PAM50 subtypes. Specifically, when the number of clusters is specified to be 2 in HCLUST, one cluster of two patients appears (this cluster does not disappear when different number of clusters are specified), suggesting that these two patients are very different from the majority of breast cancer patients. In fact, these two patients are the only breast cancer patients who are classified into Glioblastoma patients based on the RNAseq expression profile of the 6000 most variable genes in the Pan-cancer analysis as done in Hoadley et al. (2014). Therefore, these two patients might represent an interesting subtype. As suggested by the posterior similarity matrix shown in Figure 2B, there might be more than 10 clusters in the breast cancer dataset and, in fact, the MPEAR method yields 36 clusters with some clustering having a few patients. Detailed investigation on these clusters by integrating other

information such as the copy number variation profiles might provide hints on some clinically important subtypes.

5.2 Lung disease data

The second dataset is the gene expression data for 474 patients with lung diseases from the Lung Genomics Research Consortium (LGRC). The 474 patients are broadly divided into two groups with 255 patients having interstitial lung disease (ILD) (Schwarz and King, 2003) and 219 patients having chronic obstructive pulmonary disease (COPD) (Vestbo, 2014) with four outlier COPD patients being removed. It is worth noting that different subgroups can exist for each group. ILD is a general category including different types of lung diseases, such as interstitial pneumonia, idiopathic pulmonary fibrosis, and others (Schwarz and King, 2003), while different subtypes have been observed in COPD (Garcia-Aymerich et al., 2011). Here, the ILD group and COPD group are treated as the true clusters, although we know there are likely more than two clusters in this dataset.

The normalized gene expression data are downloaded from the LGRC genomics website (<http://www.lung-genomics.org>) and the 482 genes with the largest variances (standard deviation greater than 1.5) are selected for clustering. The number of factors given by the parallel analysis on this dataset is 14. Similar to what we have done in the breast cancer dataset, we compare the clustering performance of these methods when we vary the number of clusters from 2 to 10. From Figure 3A, we can see that the adjusted Rand index for BCSUB is much higher than the other methods. Also, we take a closer look at the clusters identified by BCSUB to see if any interesting clusters are discovered. When the number of clusters is specified to be 3, one cluster of 19 patients appears (this cluster is robust to the increased number of clusters specified), apart from the two major clusters with the majority being ILD patients and COPD patients, respectively. In fact, 47.4% of the 19 patients are very severe COPD patients, which is significantly higher than that in the two major clusters (1.3% and 15.1%). Also, the proportion of patients with evenly distributed emphysema, one important clinical variable that guides surgical procedures for COPD patients (Boutou et al., 2015), in the 19 patients (57.9%) is higher than that in the other two clusters (5.1% and 31.2%). These suggest that the clusters identified by BCSUB is clinically relevant. And, the number of clusters given by BCSUB (MPEAR) is large (30 clusters), as also suggested in Figure 3B. Detailed investigation on these clusters is needed to confirm if these clusters represent clinically important subtypes.

For both real data applications, trace plot monitoring suggests that the models converged relatively quickly. Also, the average of differences of each entry across three posterior similarity matrices is calculated to further check if the clustering structures from the three chains have substantial differences. The mean, median, and standard deviation of the average differences for all upper diagonal entries (excluding the diagonal elements) are 0.028, 0.000, 0.078 for the breast cancer dataset and are 0.04, 0.000, 0.116 for the lung disease dataset, which suggests that the posterior similarity matrices and hence the clustering structures are similar across the three chains for both datasets.

5.3 Conclusion

In both real data applications, we find that BCSUB has similar or better performance than other commonly used clustering methods, provided that the number of clusters is appropriately chosen. However, the performance of the other methods is not stable, that is MCLUST works better in the breast cancer dataset whereas PAM works better in the lung disease dataset, and is not robust to the misspecification of the number of clusters. Moreover, BCSUB identifies some clinically interesting clusters that are missed by other methods. However, it seems BCSUB (MPEAR) yields a large number of clusters, which may or may not represent unknown subtypes in these patients. In practice, detailed investigation of these clusters given by BCSUB (MPEAR), together with visual inspection on the posterior similarity matrix and prior clinical knowledge, is highly recommended to determine the number of clusters and therefore the clustering structure.

6 Discussion

In this article, we have proposed a Bayesian clustering method, called BCSUB, for subtype identification based on gene expression profiles. This method is based on the construction of metagenes and clustering is done simultaneously with metagene construction under a single framework. In BCSUB, the factor analysis model is used to define the metagene and the DP prior is specified on the factor scores to induce clustering. Both simulation studies and real data application showed that BCSUB outperforms other commonly used clustering methods, especially in situations where the variables used for clustering are highly correlated so that they can be explained by a few latent factors and some of the variables are uninformative. While we focus on the applications to gene expression data in this paper, our proposed model can be easily extended for clustering other data types and hence can be applied to applications in other fields. For example, our proposed model can be integrated into the generalized linear model for clustering mixed data with both discrete and continuous data, as commonly encountered in social science (Yang and Dunson, 2010; Murray et al., 2013). Our proposed method, when coupled with other appropriate statistical models, can also be used for clustering temporally and/or spatially correlated data, such as brain fMRI data (Craddock et al., 2012).

In BCSUB, we use the parallel analysis to determine the number of factors. There are a few methods in the literature that incorporate the inference on the number of factors. However, these methods increase the computation time considerably, especially in the high dimensional setting. A computationally efficient approach with inference on the number of factors, which performs the model averaging step for models with different number of factors in a more natural way, could be an interesting research topic for future study.

All clustering methods come with their own ways of determining the optimal number of clusters. Some of them are statistically optimal and some of them are heuristic. For instance, we use MPEAR on the posterior similarity matrix of our methods to determine the number of clusters. However, these methods might not always provide clinically relevant clustering structure, although they may be statistically optimal. Therefore, our recommendation is to use the plot of posterior similarity matrix, together with clinical insights, to choose the number of clusters, which could produce clinically more relevant subtypes.

Our proposed method allows the incorporation of more structural constraints on the loading matrix Λ to further increase interpretability and efficiency. When the number of genes increases, the dimension of the loading matrix will increase drastically, in which case sparsity constraints on the loading matrix might be preferable to increase stability and efficiency of the method. Also, prior knowledge of the pathway structure could be used to generate structural constraints on the loading matrix to make the factors scores more biologically interpretable.

Acknowledgments

We thank the editor and reviewers for careful reading of our paper and for their insightful and constructive comments, which have greatly helped improve our work. Jiehuan Sun and Hongyu Zhao were supported by National Science Foundation grant DMS-1106738 and National Institutes of Health grants R01 GM59507 and P01 CA154295. Joshua L. Warren was supported by CTSA Grant Number UL1 TR001863 from the National Center for Advancing Translational Science (NCATS), a component of the National Institutes of Health (NIH). Its contents are solely the responsibility of the authors and do not necessarily represent the official view of NIH.

Appendix

Here, we provide details of the MCMC sampling algorithm used for fitting our proposed method. Let e_i be the cluster membership indicator, i.e. $e_i = k$ means subject i belongs to cluster k . Let K be the largest possible cluster indicator in the most recent step. Let Θ be the set of all parameters. Based on the model specification and the prior distributions given in the main text, we can write out the joint distribution of all parameters given data as follows.

$$P(\Theta | Y) = \prod_{i=1}^n \frac{\exp\left(-\frac{1}{2}(\mathbf{Y}_i - \Lambda \boldsymbol{\eta}_i)^T \Sigma^{-1} (\mathbf{Y}_i - \Lambda \boldsymbol{\eta}_i)\right)}{\sqrt{(2\pi)^G |\Sigma|}} \times \prod_{i=1}^n \prod_{k=1}^K \left\{ \frac{\exp\left(-\frac{1}{2}(\boldsymbol{\eta}_i - \boldsymbol{\mu}_k)^T \Omega^{-1} (\boldsymbol{\eta}_i - \boldsymbol{\mu}_k)\right)}{\sqrt{(2\pi)^M |\Omega|}} \right\}^{\delta(e_i=k)} \times \prod_{k=1}^K \frac{\exp\left(-\frac{1}{2} \boldsymbol{\mu}_k^T (\rho \mathbf{I}_M)^{-1} \boldsymbol{\mu}_k\right)}{\sqrt{(2\pi)^G |\rho \mathbf{I}_M|}} \delta(\rho \in [0, 2]) \times \prod_{g=1}^G \frac{V_1^{V_2}}{\Gamma(V_1)} [\Sigma]_{gg}^{-V_1-1} \exp\left(-\frac{V_2}{[\Sigma]_{gg}}\right) \times \prod_{m=1}^M \frac{V_1^{V_2}}{\Gamma(V_1)} [\Omega]_{mm}^{-V_1-1} \exp\left(-\frac{V_2}{[\Omega]_{mm}}\right) \times \prod_{g=1}^G \prod_{m < \min\{M, g\}} \frac{\exp\left(-\frac{1}{2\sigma^2} ([\Lambda]_{gm})^2\right)}{\sqrt{2\pi\sigma^2}} \times \frac{\Gamma(c)}{\Gamma(c)+n} c^K \prod_{k=1}^K \Gamma(n_k),$$

where $c = 1$ is the concentration parameter in the DP prior, $v_1 = v_2 = 0.01$ are the parameters in the Inverse Gamma prior distributions for the diagonal elements of Σ and Ω ,

$n_k = \sum_{i=1}^n \delta(e_i=k)$ is the number of subjects in cluster k in the current step, and $[\cdot]_{gm}$ denotes the element in the g^{th} row and m^{th} column of the matrix. Then, MCMC sampling proceeds in the following steps:

1. The gene specific variances are sampled as follows.

$$[\Sigma]_{\text{gg}} | \dots \sim \text{Inverse Gamma} \left(v_1 + \frac{n}{2}, v_2 + \frac{1}{2} \sum_i^n [(\mathbf{Y}_i - \mathbf{\Lambda} \boldsymbol{\eta}_i)(\mathbf{Y}_i - \mathbf{\Lambda} \boldsymbol{\eta}_i)^T]_{\text{gg}} \right),$$

where $|\dots$ means conditional all the other parameters and data.

2. The diagonal elements of the covariance matrix $\mathbf{\Omega}$ are sampled as follows.

$$[\mathbf{\Omega}]_{\text{mm}} | \dots \sim \text{Inverse Gamma} \left(v_1 + \frac{n}{2}, v_2^* \right),$$

where $v_2^* = v_2 + \frac{1}{2} \sum_i \sum_{k=1}^K \delta(e_i=k) [(\boldsymbol{\eta}_i - \boldsymbol{\mu}_k)(\boldsymbol{\eta}_i - \boldsymbol{\mu}_k)^T]_{\text{mm}}$.

3. The subject specific factor scores $\boldsymbol{\eta}_i$ are sampled as follows.

$$\boldsymbol{\eta}_i | \dots \sim \text{MVN} \left(\Omega^* \left(\mathbf{\Lambda}^T \Sigma^{-1} \mathbf{Y}_i + \Omega^{-1} \boldsymbol{\mu}_k \right), \Omega^* \right),$$

where $\Omega^* = (\mathbf{\Lambda}^T \Sigma^{-1} \mathbf{\Lambda} + \Omega^{-1})^{-1}$.

4. The loading matrix $\mathbf{\Lambda}$ are sampled as follows.

$$[\mathbf{\Lambda}]_{g \cdot} | \dots \sim \text{MVN} \left(\Omega_g^* \left([\Sigma]_{\text{gg}}^{-1} \boldsymbol{\eta}^T [\mathbf{Y}]_{\cdot g} \right), \Omega_g^* \right),$$

where $\Omega_g^* = \left([\Sigma]_{\text{gg}}^{-1} \boldsymbol{\eta}^T \boldsymbol{\eta} + \sigma^{-2} \mathbf{I}_M \right)^{-1}$, σ^2 is the prior variance for elements in $\mathbf{\Lambda}$, $\boldsymbol{\eta}^T = [\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_n]$, and $[\mathbf{Y}]_{\cdot g}$ is the g^{th} column of $\mathbf{Y} = [\mathbf{Y}_1, \dots, \mathbf{Y}_n]^T$. For constrained rows of $\mathbf{\Lambda}$, where we only need to update m elements, the updating rule is similar except that $\boldsymbol{\eta}$ is constrained to the first m columns and the contribution of the $(m+1)^{\text{th}}$ column of $\boldsymbol{\eta}$ is deducted from the corresponding column of $[\mathbf{Y}]$.

5. The cluster specific means $\boldsymbol{\mu}_k$ are sampled as follows.

$$\boldsymbol{\mu}_k | \dots \sim \text{MVN} \left(\Omega_k^* \left(\mathbf{\Omega}^{-1} \sum_{i=1}^n \boldsymbol{\eta}_i \delta(e_i=k) \right), \Omega_k^* \right),$$

where $\Omega_k^* = (n_k \mathbf{\Omega}^{-1} + (\rho \mathbf{I}_M)^{-1})^{-1}$.

6. The cluster membership indicator for each subject is sampled as follows. For cluster k , which is occupied by some subjects excluding subject i , we have

$$P_k := P(e_i = k | e_{-i}, \mathbf{\Omega}, \rho, \eta) = l \times n_k^{(-i)} \times \Phi\left(\boldsymbol{\eta}_i; \tilde{\boldsymbol{\mu}}_k^{(-i)}, \tilde{\boldsymbol{\Omega}}_k^{(-i)}\right)$$

where l is some positive constant shared across all clusters, $\Phi(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the density function for Multivariate Normal distribution with mean $\boldsymbol{\mu}$ and

covariance matrix $\boldsymbol{\Sigma}$, $n_k^{(-i)} = \sum_{j=1, j \neq i}^n \delta(e_j = k)$ is the number of subjects in cluster k excluding i^{th} subject, and $\tilde{\boldsymbol{\mu}}_k^{(-i)}$ and $\tilde{\boldsymbol{\Omega}}_k^{(-i)}$ are the cluster specific means and variances for the factor scores respectively, which can be calculated as

$$\begin{aligned}\tilde{\boldsymbol{\Omega}}_k^{(-i)} &= \left(n_k^{(-1)} \boldsymbol{\Omega}^{-1} + (\rho \mathbf{I}_M)^{-1}\right)^{-1} + \boldsymbol{\Omega} \\ \tilde{\boldsymbol{\mu}}_k^{(-i)} &= \left(n_k^{(-1)} \boldsymbol{\Omega}^{-1} + (\rho \mathbf{I}_M)^{-1}\right)^{-1} \left(n_k^{(-1)} \boldsymbol{\Omega}^{-1} \bar{\boldsymbol{\eta}}_k^{(-i)}\right) \\ \bar{\boldsymbol{\eta}}_k^{(-i)} &= \frac{\sum_{j=1, j \neq i}^n \boldsymbol{\eta}_j \delta(e_j = k)}{n_k^{(-1)}}.\end{aligned}$$

For a new cluster $K + 1$, we have

$$P_{K+1} := P(e_i = K + 1 | e_{-i}, \mathbf{\Omega}, \rho, \eta) = l \times c \times \Phi(\boldsymbol{\eta}_i; \mathbf{0}_M, \rho \mathbf{I}_M + \boldsymbol{\Omega}).$$

Then, the cluster membership indicator is drawn from a multinomial distribution, that is

$$e_i | e_{-i}, \mathbf{\Omega}, \rho, \eta \sim \text{multinomial}(P_1, \dots, P_{K+1}).$$

7. The variance parameter ρ in the base distribution of the DP prior can be sampled as follows. For convenience, we first transform the variance parameter to avoid the positive constraint on the variance parameter and the posterior sampling is

conducted on the transformed variance parameter. Specifically, let $\tau = \log\left(\frac{a - \rho}{\rho - b}\right)$, where a and b are the parameters in the prior distribution for $\rho \sim \text{Uniform}(a, b)$ ($a = 0$ and $b = 2$ in our case). Then, a Metropolis Hasting updating step can be performed to draw τ using $N(\tau', \omega^2)$ as the proposal distribution, where τ' is the current value and ω^2 can be used to tune the acceptance rate. The acceptance probability can be calculated as

$$A(\tau | \tau') = \frac{P(\tau) P(\{\boldsymbol{\mu}_k\}_{k=1}^K | \tau)}{P(\tau') P(\{\boldsymbol{\mu}_k\}_{k=1}^K | \tau')},$$

$$\text{where } P(\tau) = \frac{\exp(\tau)}{(\exp(\tau) + 1)^2} \text{ and}$$

$$P(\{\boldsymbol{\mu}_k\}_{k=1}^K | \tau) \propto \prod_{k=1}^K \Phi \left(\boldsymbol{\mu}_k; \mathbf{0}_M, \left(\frac{\exp(\tau) + a}{\exp(\tau) + 1} \right) \mathbf{I}_M \right).$$

References

- Binder DA. Bayesian cluster analysis. *Biometrika*. 1978; 65:31–38.
- Boutou AK, Zoumot Z, Nair A, Davey C, Hansell DM, Jamurtas A, Polkey MI, Hopkinson NS. The impact of homogeneous versus heterogeneous emphysema on dynamic hyperinflation in patients with severe COPD assessed for lung volume reduction. *COPD J Chronic Obstr Pulm Dis*. 2015; 12:598–605.
- Cali ski T, Harabasz J. A dendrite method for cluster analysis. *Commun Stat*. 1974; 3:1–27.
- Carvalho CM, Chang J, Lucas JE, Nevins JR, Wang Q, West M. High-dimensional sparse factor modeling: applications in gene expression genomics. *J Am Stat Assoc*. 2008; 103:1438–1456. [PubMed: 21218139]
- Craddock RC, James GA, Holtzheimer PE, Hu XP, Mayberg HS. A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Hum Brain Mapp*. 2012; 33:1914–1928. [PubMed: 21769991]
- Dahl, DB. *Bayesian Inference for Gene Expression and Proteomics*. Cambridge: Cambridge University Press; 2006. Model-based clustering for expression data via a Dirichlet process mixture model; p. 201–218.
- Drasgow F, Lissak RI. Modified parallel analysis: a procedure for examining the latent dimensionality of dichotomously scored item responses. *J Appl Psychol*. 1983; 68:363–373.
- Erosheva, EA., Curtis, SM. Technical Report 589. University of Washington; 2011. Dealing with rotational invariance in Bayesian confirmatory factor analysis.
- Ferguson TS. A Bayesian analysis of some nonparametric problems. *Ann Stat*. 1973; 1:209–230.
- Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. *J Am Stat Assoc*. 2002; 97:611–631.
- Fritsch A, Ickstadt K. Improved criteria for clustering based on the posterior similarity matrix. *Bayesian Anal*. 2009; 4:367–391.
- Garcia-Aymerich J, Gómez FP, Benet M, Farrero E, Basagaña X, Gayete À, Paré C, Freixa X, Ferrer J, Ferrer A, Roca J. Identification and prospective validation of clinically relevant chronic obstructive pulmonary disease (COPD) subtypes. *Thorax*. 2011; 66:430–437. [PubMed: 21177668]
- Geweke J, Zhou G. Measuring the pricing error of the arbitrage pricing theory. *Rev Financ Stud*. 1996; 9:557–587.
- Hartigan JA, Wong MA. Algorithm AS 136: a k-means clustering algorithm. *J R Stat Soc Ser C Appl Stat*. 1979; 28:100–108.
- Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, Leiserson MD, Niu B, McLellan MD, Uzunangelov V, Zhang J. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*. 2014; 158:929–944. [PubMed: 25109877]
- Hoyle, RH., Duvall, JL. Determining the number of factors in exploratory and confirmatory factor analysis. In: Kaplan, D., editor. *Handbook of Quantitative Methodology for the Social Sciences*, chapter 16. Thousand Oaks, CA: Sage; 2004. p. 301–315.
- Hubert L, Arabie P. Comparing partitions. *J Classif*. 1985; 2:193–218.
- Jain S, Neal RM. A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *J Comput Graph Stat*. 2004; 13:158–182.
- Jasra A, Holmes C, Stephens D. Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Stat Sci*. 2005; 20:50–67.
- Jeste SS, Geschwind DH. Disentangling the heterogeneity of autism spectrum disorder through genetic findings. *Nat Rev Neurol*. 2014; 10:74–81. [PubMed: 24468882]
- Johnson SC. Hierarchical clustering schemes. *Psychometrika*. 1967; 32:241–254. [PubMed: 5234703]

- Kalli M, Griffin JE, Walker SG. Slice sampling mixture models. *Stat Comput.* 2011; 21:93–105.
- Kim S, Tadesse MG, Vannucci M. Variable selection in clustering via Dirichlet process mixture models. *Biometrika.* 2006; 93:877–893.
- Lee, DD., Seung, HS. *Advances in Neural Information Processing Systems*. Vol. 13. Boston, MA, USA: MIT Press; 2001. Algorithms for non-negative matrix factorization; p. 556–562.
- Liu W, Yuan K, Ye D. Reducing microarray data via nonnegative matrix factorization for visualization and clustering analysis. *J Biomed Inform.* 2008; 41:602–606. [PubMed: 18234564]
- Lock EF, Dunson DB. Bayesian consensus clustering. *Bioinformatics.* 2013; 29:2610–2616. [PubMed: 23990412]
- Lopes HF, West M. Bayesian model assessment in factor analysis. *Stat Sin.* 2004; 14:41–67.
- MacEachern SN. Estimating normal means with a conjugate style Dirichlet process prior. *Commun Stat Simul Comput.* 1994; 23:727–741.
- Medvedovic M, Yeung KY, Bumgarner RE. Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics.* 2004; 20:1222–1232. [PubMed: 14871871]
- Murray JS, Dunson DB, Carin L, Lucas JE. Bayesian Gaussian copula factor models for mixed data. *J Am Stat Assoc.* 2013; 108:656–665. [PubMed: 23990691]
- Neal, RM. *Maximum Entropy and Bayesian Methods*. Berlin: Springer; 1992. Bayesian mixture modeling; p. 197–211.
- Pan W, Shen X. Penalized model-based clustering with application to variable selection. *J Mach Learn Res.* 2007; 8:1145–1164.
- Papaspiliopoulos O, Roberts GO. Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika.* 2008; 95:169–186.
- Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, Quackenbush JF. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol.* 2009; 27:1160–1167. [PubMed: 19204204]
- Perou CM, Sørli T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge Ø. Molecular portraits of human breast tumours. *Nature.* 2000; 406:747–752. [PubMed: 10963602]
- Qin ZS. Clustering microarray gene expression data using weighted Chinese restaurant process. *Bioinformatics.* 2006; 22:1988–1997. [PubMed: 16766561]
- Rodriguez A, Dunson DB. Functional clustering in nested designs: modeling variability in reproductive epidemiology studies. *Ann Appl Stat.* 2014; 8:1416–1442.
- Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math.* 1987; 20:53–65.
- Schwarz, MI., King, TE. *Interstitial lung disease*. 5th. Shelton, CT: People's Medical Publishing House-USA; 2003.
- Sethuraman J. A constructive definition of Dirichlet priors. *Stat Sin.* 1994; 4:639–650.
- Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen M, Van de Rijn M, Jeffrey S, Thorsen T. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci.* 2001; 98:10869–10874. [PubMed: 11553815]
- Tamayo P, Scanfeld D, Ebert BL, Gillette MA, Roberts CW, Mesirov JP. Metagene projection for cross-platform, cross-species characterization of global transcriptional states. *Proc Natl Acad Sci.* 2007; 104:5959–5964. [PubMed: 17389406]
- The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature.* 2012; 490:61–70. [PubMed: 23000897]
- Vestbo J. COPD: definition and phenotypes. *Clin Chest Med.* 2014; 35:1–6. [PubMed: 24507832]
- Walker SG. Sampling the Dirichlet mixture model with slices. *Commun Stat Simul Comput.* 2007; 36:45–54.
- West, M. Bayesian factor regression models in the “large p, small n” paradigm,”. In: Bernardo, J.Bayarri, M.Dawid, A.Heckerman, D.Smith, A., West, M., editors. *Bayesian Statistics*. Vol. 7. Oxford: Oxford University Press; 2003. p. 723–732.

- Wigle DA, Jurisica I, Radulovich N, Pintilie M, Rossant J, Liu N, Lu C, Woodgett J, Seiden I, Johnston M, Keshavjee S. Molecular profiling of non-small cell lung cancer and correlation with disease-free survival. *Cancer Res.* 2002; 62:3005–3008. [PubMed: 12036904]
- Wold S, Esbensen K, Geladi P. Principal component analysis. *Chemometr Intell Lab Syst.* 1987; 2(1–3):37–52.
- Yang M, Dunson DB. Bayesian semiparametric structural equation models with latent variables. *Psychometrika.* 2010; 75:675–693.
- Yeung KY, Ruzzo WL. Principal component analysis for clustering gene expression data. *Bioinformatics.* 2001; 17:763–774. [PubMed: 11590094]

Figure 1.

(A) The adjusted Rand index for all the methods in the breast cancer dataset, where the true subtypes defined by PAM50. (B) The adjusted Rand index for all the methods in the breast cancer dataset, where the true subtypes defined by ER.

Figure 2.

(A) The adjusted Rand index for all the methods in the breast cancer dataset, where the true subtypes defined by PR. (B) The posterior similarity matrix in the breast cancer dataset.

Figure 3.
 (A) The adjusted Rand index for all the methods in the lung disease dataset. (B) The posterior similarity matrix in the lung disease dataset.

Table 1

Comparisons of the five clustering methods in all scenarios under different settings.

Scenarios	BCSub	MCLUST	PAM	PCA	NMF
F0	0.94 (0.23)	0.05 (0.05)	0.10 (0.09)	0.96 (0.09)	0.02 (0.02)
F50	0.98 (0.14)	0.02 (0.02)	0.06 (0.08)	0.92 (0.17)	0.01 (0.02)
F100	1.00 (0.00)	0.01 (0.02)	0.02 (0.03)	0.86 (0.28)	0.01 (0.02)
R0	0.92 (0.24)	0.53 (0.06)	0.88 (0.10)	0.41 (0.48)	0.18 (0.16)
R50	0.90 (0.27)	0.54 (0.06)	0.77 (0.15)	0.04 (0.20)	0.29 (0.19)
R100	0.85 (0.33)	0.49 (0.08)	0.43 (0.23)	0.00 (0.00)	0.29 (0.15)

The numbers in each cell indicate the adjusted Rand index for each method under each scenario with standard deviation in the parenthesis. The bold numbers indicate the top performers for each scenario (The methods having differences in the average adjusted Rand index less than 0.02 are considered to have similar performance).