Published in final edited form as:

J Am Stat Assoc. 2013 January 1; 108(503): . doi:10.1080/01621459.2013.784705.

A Nonparametric Bayesian Model for Local Clustering with Application to Proteomics

Juhee Lee,

Department of Statistics, The Ohio State University, Columbus, OH

Peter Müller,

Department of Mathematics, University of Texas Austin, Austin, TX

Yitan Zhu, and

Center for Clinical and Research Informatics, NorthShore University HealthSystem, Evanston, IL

Yuan Ji*

Center for Clinical and Research Informatics, NorthShore University HealthSystem, Evanston, IL

Abstract

We propose a nonparametric Bayesian local clustering (NoB-LoC) approach for heterogeneous data. NoB-LoC implements inference for nested clusters as posterior inference under a Bayesian model. Using protein expression data as an example, the NoB-LoC model defines a protein (column) cluster as a set of proteins that give rise to the same partition of the samples (rows). In other words, the sample partitions are nested within protein clusters. The common clustering of the samples gives meaning to the protein clusters. Any pair of samples might belong to the same cluster for one protein set but to different clusters for another protein set. These local features are different from features obtained by global clustering approaches such as hierarchical clustering, which create only one partition of samples that applies for all the proteins in the data set. In addition, the NoB-LoC model is different from most other local or nested clustering methods, which define clusters based on common parameters in the sampling model. As an added and important feature, the NoB-LoC method probabilistically excludes sets of irrelevant proteins and samples that do not meaningfully co-cluster with other proteins and samples, thus improving the inference on the clustering of the remaining proteins and samples. Inference is guided by a joint probability model for all the random elements. We provide a simulation study and a motivating example to demonstrate the unique features of the NoB-LoC model.

Keywords

Dirichlet Process; Pólya Urn; Protein Expression; RPPA; Random Partitions

1 Introduction

1.1 Motivation

Clustering methods have been widely applied to heterogeneous data to investigate groups of measurements with similar patterns. The identification of such groups can provide scientific insight. For example, in biological research, clustering of data is widely used to study functionally similar groups of molecules (such as proteins and genes) or similar biological

*Address for Correspondence: Center for Clinical and Research Informatics, NorthShore University HealthSystem, Evanston, IL 60201 USA. yji@northshore.org.

Supplementary materials: Supplementary materials are available under the Paper Information link at the JASA website.

tissues and samples that can help researchers understand networks of molecules and subtypes of samples (Belacel *et al.*, 2006). Despite the large number of existing clustering methods, researchers continue to improve clustering solutions to meet new challenges that arise from new experimental platforms and new data formats. In the following discussion we assume that the data is generated from a proteomics experiment, and is arranged as a matrix with rows corresponding to experimental units (samples) and columns corresponding to features (proteins). To simplify the discussion and in anticipation of the motivating application we refer to rows and columns as samples and proteins, respectively.

An important and open problem is to allow different partitions of the samples for different subsets of proteins. There may be multiple subsets of proteins that define different biological processes, resulting in multiple ways of clusterings the same set of samples. Another desirable feature in many applications is the exclusion of a subset of columns and/or rows from the cluster arrangement. Scientific discoveries have supported the need for these features. For example, previous studies have demonstrated that most biological processes involve only a small subset of biological features, and different biological processes partition biological samples differently (Jiang *et al.*, 2004). In other words, only a subset of proteins might be informative about the partition of samples of interest, and the remaining proteins are irrelevant to the sample clustering. Similarly, some samples whose underlying disease mechanism does not involve a given protein set may not meaningfully co-cluster with other samples with respect to these proteins.

1.2 RPPA Data

The proposed method is motivated by a data set from an experiment using reverse phase protein arrays (RPPA). RPPA is a new functional proteomics interrogation platform to study molecular profiling of disease subtypes based on protein expression (Tibes *et al.*, 2006). It allows to simultaneously quantify the expression of many proteins in a biological network for a large number of patients. The data set records expression levels of $G = 55$ proteins that were selected from two cell signaling pathways (PI3K and MAPK) for $N = 256$ samples from breast cancer patients. The heatmap of the data set is shown in Figure 1 along with dendrograms from hierarchical clustering.

The RPPA data set includes three important biomarker proteins: ER, PR and HER2 (# 53, 54 and 55 in the figure). Clinically three subtypes of breast cancer, namely HR+, HER2+ and TN are based on these three biomarker proteins. For samples with subtype HR+, either ER or PR is over-expressed, but HER2 is under-expressed. For samples with subtype HER2+, HER2 is over-expressed regardless of the expression of ER or PR. For samples with subtype TN standing for triple negative, all three biomarker proteins are under-expressed. The classification of breast cancer into these three subtypes is often used to recommend different treatment options. However, recent studies have identified new standards and markers for defining clinically meaningful subtypes of breast cancer. For example, Lehmann *et al.* (2011) discovered six stable groups within the TN breast cancer for more appropriate targeted treatments. In two recent Nature publications Curtis *et al.* (2012) and The Cancer Genome Atlas Network (2012) suggest that breast cancer encompasses a variety of different subdiseases based on genomic and transcriptomic profiling.

1.3 Nested Partitions

The main inference goal in the motivating RPPA study is to develop a meaningful and comprehensive profiling of samples for possibly more accurate disease prognosis. To achieve this goal we propose a nonparametric Bayesian local clustering (NoB-LoC) approach that defines a random bidirectional nested clustering on subsets of the data matrix.

The NoB-LoC model identifies protein sets that characterize different biologic processes. Central to the construction is the notion that a protein set that corresponds to a particular biologic process should give rise to a partition of samples into subsets with respect to that biologic process. For example, if a protein set were to form a particular pathway, then it might partition the samples into subgroups that correspond to different disruptions of that pathway, in the simplest case over- vs. under-activation of the pathway, as well as perhaps a remaining set without any related pathology. We formalize this idea by two features of the proposed approach.

First, we define protein sets through common (nested) partitions of the samples. That is, the NoB-LoC model assigns any two proteins into the same protein set if the two proteins partition samples in the same way, regardless of the actually observed values. This characterization of protein sets is in contrast to a traditional characterization by means of some common cluster-specific parameters in the sampling model, usually a location parameter. As a result, the NoB-LoC model produces sample partitions that are nested within protein sets, allowing the same samples to be clustered differently for different protein sets. Clusters of samples are only meaningful with respect to a particular protein set. The corresponding submatrix involving those samples and proteins is referred to as a “local cluster”.

In addition, the NoB-LoC method probabilistically excludes some proteins and samples from clustering and partitions only subsets of the data. In many applications the data include some proteins and samples that are irrelevant to any meaningful clustering. These irrelevant proteins and/or samples may introduce extra noise that complicates the identification of clustering patterns in the remaining data. The additional noise reduces the power of finding the signals of interest. The NoB-LoC method explicitly models such irrelevant proteins by allowing a special “inactive” protein set that does not relate to any clustering of samples. For each protein set that does induce sample clusters the NoB-LoC model also includes a special cluster of “inactive” samples that do not co-cluster with any other samples.

Figure 2 illustrates these features with a stylized data matrix, arranged according to the cluster membership of proteins and samples. Protein sets 1 and 2 are active and protein set 0 is inactive. Cells in white represent inactive samples (such as sample cluster 0 in protein set 2). The remaining cells with matching colors in the same column form sample clusters. Distinct colors (grey shades) indicate distinct values of parameters. For example, proteins {P2, P4, P6} in protein set 2 all exhibit the same three sample clusters: Sample cluster 0 is inactive, including samples {S2, S4, S7, S8}. Each of the two active sample clusters, {S3, S5, S9} and {S1, S6}, forms a local cluster. Within each local cluster, the colors are the same across the samples but vary across proteins, because protein sets are defined by matching partitions of samples across all proteins rather than by matching means or other parameters. This highlights a key difference between clustering proteins by means vs. clustering by (nested) sample partitions. Most existing clustering approaches define clusters by sharing some parameters, typically location or scale parameters, across all items in a cluster so that the row clusters are represented by the same color across columns, e.g. samples in sample cluster 1 of protein set 2 would all have the same color for proteins {P2, P4, P6}. As a result, posterior inference would include shrinkage across proteins. This is not appropriate for the motivating application, and it is not assumed in the proposed NoB-LoC scheme.

In Section 2 we will introduce the model starting with a partition of the proteins. Alternatively the local partition of the NoB-LoC model could be described with a focus on samples and the G —dimensional response (row) vectors. The proposed model boils down to a partition of the G —dimensional response vector that changes as we move across samples.

This can be seen in Figure 2. Partitioning subset (of proteins) are consistent only across all samples in a sample cluster relative to that protein set. This alternative view also highlights the asymmetric nature of the model.

1.4 Current Approaches and Limitations

There is an extensive literature on clustering methods for statistical inference. Among the most widely used approaches are algorithmic methods such as K-means and hierarchical clustering. Other methods are based on probability models, including the popular model-based clustering. For a review, see Fraley and Raftery (2002). A special type of model-based clustering methods includes approaches that are based on nonparametric Bayesian inference (Quintana, 2006). The idea of these approaches is to construct a discrete random probability measure and use the arrangement of ties that arise in random sampling from a discrete distribution to define random clusters. Rather than fixing the number of clusters, nonparametric Bayesian models naturally imply a random number and size of clusters. For example, the Dirichlet process prior, which is arguably the most commonly used nonparametric Bayesian model, implies infinitely many clusters in the population, and an unknown, but finite number of clusters for the observed data. Recent examples of nonparametric Bayesian clustering have been described in Medvedovic and Sivaganesan (2002), Dahl (2006), and Müller *et al.* (2011) among others.

Recall that we use “proteins” to refer to the columns and “samples” to refer to the rows in a data matrix. The methods described above are one-dimensional clustering methods that yield a single partition of all samples that applies across all proteins (or vice versa). We refer these methods as “global clustering methods” in the subsequent discussion. In contrast to global clustering methods, local clustering methods are bidirectional and aim at discovering local patterns involving only subsets of proteins and/or samples. This requires simultaneous clustering of proteins and samples in a data matrix. The basic concept of local clustering has been described in Cheng and Church (2000). Many authors proposed nonparametric Bayesian approaches for local clustering. These include Meeds and Roweis (2007), Dunson (2009), Petrone *et al.* (2009), Rodríguez *et al.* (2008), Dunson *et al.* (2008), Roy and Teh (2009), Wade *et al.* (2011) and Rodríguez and Ghosh (2012). Except for the nested infinite relational model of Rodríguez and Ghosh (2012) these methods do not explicitly define a sample partition that is nested within protein sets and some of the methods need tweaking to be used as a prior model for clustering of samples and proteins in our data matrix. For example, the enriched Dirichlet process (Wade *et al.*, 2011) implies a discrete random probability measure P for $x_g \sim P$ and for each unique value x^* among the x_g a discrete random probability measure Q_{x^*} . We could interpret the x_g as protein-specific labels and use them to define a random partition of proteins (the x_g 's have no further use beyond inducing the partition of proteins). Using protein set 2 in Figure 2 for an illustration, $x_1 = x_3 = x_1^*$, $x_2 = x_4 = x_6 = x_2^*$ and $x_5 = x_0^*$ defines three protein sets. The random distributions $Q_{x_s^*}$ can then be used to generate sample/protein-specific parameters, $\theta_{ig} | x_g = x_s^* \sim Q_{x_s^*}$, $s = 1, \dots, S$, and ties among the θ_{ig} can be used to define sample clusters. The resulting prior on sample clusters is identical to the prior under the nested Dirichlet process (Rodríguez *et al.*, 2008), although the two models arise from very different constructions and motivations. An important observation is that under both models only the prior $Q_{x_s^*}$ is shared for all proteins in the same protein cluster. The actual sample clusters are independent realizations of the random (sample) partition that is implied by $\theta_{ig} \sim Q_{x_s^*}$. For example, referring again to the stylized example in Figure 2, one could get $_{32} = _{52} = _{92}$ to define a sample cluster $\{3, 5, 9\}$ under protein $g = 2$, but differently for $g = 4$, as only the prior $Q_{x_s^*}$ is shared for $_{12}$ and $_{14}$. Similar issues arise with the other models listed above. All methods above, including the nested infinite relational model (Rodríguez and Ghosh, 2012), define clusters by matching parameters, i.e., they can be characterized as non-parametric model-based clustering. The

implication is that, for example, the colors in Figure 2 would be the same across proteins as well as across samples within local clusters under the nested infinite relational model. In contrast, the NoB-LoC model focuses on partitions and allows samples to have different parameters for proteins in a protein set. In addition, the NoB-LoC model allows inactive proteins and samples and partitions subsets of proteins and samples.

We review in some more detail three existing methods that are most relevant to the proposed approach.

DCIM—Freudenberg *et al.* (2010) developed the differential co-expression infinite mixture (DCIM) model for gene expression data. The original description of the DCIM defines nested partitions of genes, nested within subsets (“contexts”) of samples. But the model can alternatively be used for the desired nested clustering of samples within protein sets by simple transposition of the data matrix. To construct a nested clustering as in Figure 2 the method would define a global clustering of all samples and then define a “context” as a group of proteins that locally modifies the global clustering in the same way by combining some global clusters into one (larger) context-specific cluster. In other words, the global clustering of samples is the intersection of all local partitions of samples under the contexts. This method uses Dirichlet process priors as probability models for the involved random partitions.

Plaid Model—Lazzeroni and Owen (2002) developed the “plaid models” which describe gene expression data as a sum of biclusters (the authors call them layers). Each bicluster in their model is defined as a group of genes and samples, such that the genes are expressed similarly within the given set of samples. The plaid models then let the genes and samples in a bicluster share a common parameter, which could indicate the presence of a particular biological process. They used the *EM* algorithm to search for biclusters. Turner *et al.* (2005) presented an improved algorithm for fitting the plaid model using the example of a microarray data analysis. They sequentially searched for a cluster of genes and a corresponding cluster of samples that together formed a bicluster. The algorithm includes a stopping rule such as a maximum number of biclusters. However, the methods do not include a probabilistic description of the uncertainty in the identified bicluster.

Sparse Clustering—Recently Witten and Tibshirani (2010) developed methods of clustering samples based on a selected subset of genes for gene expression data. Their method adaptively selects genes and applies the hierarchical or K-means clustering to cluster samples based on the chosen genes. The method uses data from all samples to construct sample clusters.

Finally, we mention an alternative approach to the motivating problem of grouping samples differently with respect to different subsets of proteins or genes. Carvalho *et al.* (2008) use sparse factor analysis to model gene expression data for breast cancer patients similar to our motivating application. While the model does not enforce a partition of genes and samples with non-overlapping subsets, the inference aim is similar to ours. They report groupings of patients relative to different subsets of genes. The desired inference is achieved by imposing a prior on a factor loading matrix that favors mostly 0 entries, allowing non-zero entries to be interpreted as defining groups of samples and genes.

Building on these earlier successes, we propose the NoB-LoC method to implement desired inference as described by Figure 2. Section 2 describes the probability model and its posterior inference. Section 3 describes a simulation study. In Section 4 we report the analysis of the data set described in Section 1.2. The last section concludes with a final discussion.

2 Model and Posterior Inference

2.1 Model

We let $\mathbf{Y}$ denote an $(N \times G)$ matrix of continuous data y_{ig} observed for sample i and protein g , $i = 1, \dots, N$, and $g = 1, \dots, G$. Conditional on protein- and sample-specific parameters θ_{ig} and σ_g^2 , we assume a normal sampling model, $y_{ig} \sim N(\theta_{ig}, \sigma_g^2)$, independently across i and g . We do not consider discrete data. If desired, the model could be easily modified to accommodate different data formats and sampling models.

We start the prior construction with a random partition of the proteins, $g = 1, \dots, G$, into protein sets. A subset of G proteins is identified as active proteins and partitioned into S subsets indexed by $s = 1, \dots, S$. Here G and S are random variables for which we define a prior model below. The remaining $(G - G)$ inactive proteins are not assigned to any of the S clusters. Formally, they could be viewed as an $(S + 1)^{\text{st}}$ protein set. However, in anticipation of the upcoming discussion, it is convenient to separate them out as a special protein set, indexed by $s = 0$, referred to as the inactive protein set.

Let $w_g \in \{0, 1, \dots, S\}$ denote a random cluster membership indicator for protein g , with $w_g = s$ if protein g is in protein set s and let $\mathbf{w} = (w_1, \dots, w_G)$. We start the prior construction $p(\mathbf{w})$ by assuming that $w_g = 0$ with probability $(1 - \pi_0)$, independently for each protein. If $w_g > 0$, then proteins are allocated into one of the S active protein sets under a Pólya urn prior. In summary, we use a zero-enriched Pólya urn scheme (Sivaganesan *et al*, 2011) to define

$$p(\mathbf{w}) = \pi_0^{G'} (1 - \pi_0)^{G - G'} \frac{\alpha^S \prod_{s=1}^S \Gamma(p_s)}{\prod_{g=1}^{G'} (\alpha + g - 1)}, \quad (1)$$

where p_s is the cardinality of protein set s and α is the total mass parameter of the Pólya urn. In words, protein g is active ($w_g > 0$) or inactive ($w_g = 0$) with probability π_0 and $(1 - \pi_0)$, respectively. Given that the protein is active, it may join a previously formed protein set with probability proportional to the size p_s of the protein sets or start a new singleton protein set by itself with probability proportional to α .

Alternative models could be substituted for $p(\mathbf{w})$ without any change to the rest of the prior model. The Pólya urn can be defined as the distribution of ties that is implied by i.i.d. sampling from a discrete probability measure with a Dirichlet process prior (Blackwell and MacQueen, 1973) (The paper refers to the Dirichlet process as Ferguson distribution). Consider $\mathbf{g} \sim F$, $g = 1, \dots, G$, with a Dirichlet process prior $F \sim \text{DP}(\alpha, G^*)$ with total mass parameter α and base measure G^* Ferguson (1973). The a.s. discrete nature of F implies a positive probability for ties among the $\mathbf{g}$. Let θ_s^* , $s = 1, \dots, S$, denote the S unique values among the $\mathbf{g}$ and define clusters as $w_g = s$ if $\theta_g = \theta_s^*$. The implied prior on $\mathbf{w}$ is exactly (1) with $\pi_0 = 0$. Note that the definition (1) makes no use of any gene-specific parameters. Random partitions implied by the Dirichlet process prior are some of the most popular prior models for clustering in the literature, simply because of computational ease and analytic tractability. However, some features of the implied Pólya urn prior $p(\mathbf{w})$ might be undesirable in some applications. In particular, the Dirichlet process prior implies uneven cluster sizes, with a priori geometrically decreasing cluster sizes. This is inappropriate in many applications. However, we find that this prior feature is usually overwhelmed by the likelihood and does not persist in posterior inference. But if desired, any alternative model $p(\mathbf{w})$ could be substituted. For example, the prior implied by a Pitman-Yor process, a generalization of the Dirichlet process, could be used.

Next, we define random partitions of the samples, nested within protein sets. That is, we construct S parallel partitions, one for each active protein set s , $s = 1, \dots, S$. This formalizes the notion that samples are partitioned differently with respect to different processes. For protein set s , $s = 1, \dots, S$, we assume that the N samples are independently partitioned into $(K_s + 1) - N$ clusters. This implies, in particular, that if two proteins are in the same active protein set, they give rise to the same nested partition of samples. It also gives meaning to active protein sets as subsets of proteins related to a common process, since the fact that two proteins induce the same partition of samples is evidence of co-regulation.

We introduce a vector $\mathbf{c}_s = (c_{si}, i = 1, \dots, N)$ of cluster allocations to define the partition of samples with respect to protein set s , $s = 1, \dots, S$. Note that c_{0j} is not defined since protein set 0 only includes inactive proteins and does not imply sample clusters. Here $c_{si} = k$ means that sample i belongs to cluster k under protein set s , where $k = 0, \dots, K_s$. Similar to $\mathbf{w}$, we include a special cluster of *inactive* samples with $c_{si} = 0$, corresponding to samples that do not exhibit any recognizable pattern with respect to proteins in protein set s . In other words, the set of inactive samples $\{i: c_{si} = 0\}$ is a combination of meaningless singleton sample clusters for protein set s . Similar to $p(\mathbf{w})$ we again use a zero-enriched Pólya urn scheme to define $p(\mathbf{c} | \mathbf{w})$,

$$p(\mathbf{c} | \mathbf{w}) = \prod_{s=1}^S p(\mathbf{c}_s) \text{ and } p(\mathbf{c}_s) = \pi_1^{N-n_{s0}} (1 - \pi_1)^{n_{s0}} \frac{M^{K_s} \prod_{k=1}^{K_s} \Gamma(n_{sk})}{\prod_{i=1}^{n_s} (M+i-1)} \quad (2)$$

where $\pi_1 = p(c_{si} > 0)$, n_{sk} is the cardinality of sample cluster k , and M is the total mass parameter of the Pólya urn. Each of the sample clusters that is induced by a protein set becomes a *local cluster* that involves the proteins in that protein set and the samples in that sample cluster. Note that (2) implies that protein sets independently partition samples. In particular, any two samples can be clustered together for one protein set, but not for another protein set.

Recall the sampling model for the observed protein expression, $y_{ig} \sim N(\theta_{ig}, \sigma_g^2)$. For a given arrangement of protein sets $\mathbf{w}$ and protein-set-specific partitions $\mathbf{c} = (\mathbf{c}_s, s = 1, \dots, S)$ we now define a prior $p(\theta_{ig} | \mathbf{c}, \mathbf{w})$. We induce dependence across samples in each local cluster (but not across proteins) by assuming $\theta_{ig} = \theta_{i'g} = \theta_{kg}^*$ for all i, i' in the same cluster, i.e., $c_{wgi} = c_{wgi'} = k$. Note that we collapse the θ_{ig} 's across samples but not across proteins. Here θ_{kg}^* is the unique values of θ_{ig} for all samples in the k -th sample cluster. We assume

$\theta_{kg}^* \stackrel{iid}{\sim} N(\mu_{0g}, \tau_{0g}^2)$. This prior applies for all active samples i and active proteins g . For active proteins and inactive samples $\{w_g > 0, c_{wgi} = 0\}$ we assume $\theta_{ig} \stackrel{iid}{\sim} N(\mu_{1g}, \tau_{1g}^2)$. For inactive proteins $\{w_g = 0\}$ there is no partition of samples and we assume $\theta_{ig} \stackrel{iid}{\sim} N(\mu_{2g}, \tau_{2g}^2)$.

We complete the model with priors for $\sigma_g^2, \tau_{0g}^2, \tau_{1g}^2$, and τ_{2g}^2 . We use an inverse gamma distribution $\sigma_g^2 \sim \text{IG}(a_g, b_g)$ independently across g . Similarly, we use inverse gamma priors, $\tau_{\ell g}^2 \sim \text{IG}(a_{\ell g}, b_{\ell g})$, independently across g , for $\ell = 0, 1, 2$.

2.2 Posterior Inference

Let $\boldsymbol{\theta}$ denote the vector of all θ_{ig} parameters, and similarly for $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$. The joint distribution of data and parameters is summarized as

$$p(\mathbf{Y}, \mathbf{w}, \mathbf{c}, \theta, \sigma^2, \tau^2) = p(\mathbf{w})p(\mathbf{c}|\mathbf{w})p(\theta|\mathbf{w}, \mathbf{c}, \tau^2)p(\mathbf{Y}|\theta, \sigma^2)p(\sigma^2)p(\tau^2).$$

We use Markov chain Monte Carlo (MCMC) simulation to implement posterior inference.

We iteratively update (i) i_g (or θ_{kg}^*); (ii) σ_g^2 ; (iii) τ_{0g}^2 , τ_{1g}^2 and τ_{2g}^2 ; (iv) $\mathbf{c}_s$; and (v) $\mathbf{w}$. Each update is implemented as an MCMC transition probability conditional on the currently imputed values of all other parameters.

When updating $\mathbf{w}$ and $\mathbf{c}_s$, we sequentially remove a protein or a sample from $\mathbf{w}$ and $\mathbf{c}_s$, respectively, and draw a new cluster label from the full conditional posterior distributions, marginalized with respect to . Also, we make use of a pseudo prior mechanism (Carlin and Chib, 1995) to construct a proposal for $\mathbf{c}_s$ when we implement a transition probability for w_g , in particular when we propose $w_g = S + 1$, i.e., the proposal places g into a new singleton cluster of its own. The problem is that considering $w_g = S + 1$ requires an assumption about the new sample clustering $\mathbf{c}_{S+1}$. We introduce a set of auxiliary variables $\mathbf{c}_g$, $g = 1, \dots, G$ with $\mathbf{c}_g = (\mathbf{c}_{gi}, i = 1, \dots, N)$. Think of $\mathbf{c}_g$ as a potential sample partition that could be used under a singleton protein set $\{g\}$. We define a model augmentation

$p(\tilde{\mathbf{c}}_g | \mathbf{y}_g, \tau_{0g}^2, \tau_{1g}^2)$ for the auxiliary variables $\mathbf{c}_g$ to be identical to what the posterior on $\mathbf{c}_s$ would be for a singleton protein set $\{g\}$ where $\mathbf{y}_g = (y_{1g}, \dots, y_{Ng})$. We then use $\mathbf{c}_g$ in the construction of a Metropolis-Hastings proposal for w_g . Specifically, when considering $w_g = S + 1$ we make a joint proposal ($w_g = S + 1$, $\mathbf{c}_{S+1} = \mathbf{c}_g$). In other words, the proposal distribution for the sample partition under the new protein set is deterministic by copying the currently imputed value $\mathbf{c}_g$.

We diagnose convergence and mixing of the described posterior MCMC simulation using trace plots and autocorrelation plots of imputed parameters. For both, the upcoming simulation example and the data analysis, we found no evidence for practical convergence problems. Computation times are reasonable. On a linux computer (Dual Quad Core Xeon 2.66, 32GB RAM) it takes less than 4 hours to run 2,000 MCMC iterations for $n = 120$ samples and $G = 20$ proteins. The analysis of the RPPA data with $G = 55$ proteins and $n = 256$ samples and 35,000 iterations takes 7.6 hours. More technical details of the posterior MCMC are listed in the supplementary materials.

MCMC simulation output provides full posterior inference on the protein sets and their corresponding sample clusters. Summarizing this posterior distribution starts with the report of a point estimate for the random partition. We face the practical problem of summarizing a distribution over partitions. Medvedovic and Sivaganesan (2002) and Medvedovic *et al.* (2004) solved this problem by reporting posterior probabilities that any two proteins are grouped together. Dahl (2006) went a step further and introduced a method to obtain a point estimate of the random clusters based on least-square distance from the matrix of posterior pairwise co-clustering probabilities. Both methods start with an association matrix, D^w , of dimension $G \times G$ whose (g, g') element, $d_{g, g'}^w = I(w_g = w_{g'})$ is an indicator of whether proteins g and g' are grouped together. Then they compute a matrix Q of posterior pairwise co-clustering probabilities as the posterior mean of $d_{g, g'}^w$. To account for the inactive protein set in our model, we let $d_{g, g}^w = 0$ for all g if $w_g = 0$. Following Dahl (2006), we use the least-squares summary $\mathbf{w}^{LS}$ defined as

$$\mathbf{w}^{LS} = \arg \min_{\mathbf{w}} \|D^w - Q\|^2.$$

Conditional on each protein set implied by $\mathbf{w}^{LS}$, we extract the MCMC samples of $\mathbf{c}$ corresponding to the protein set and compute the least-squares estimate $\mathbf{c}_s^{LS}$, $s = 1, \dots, S^{LS}$ as the sample clusters. Alternatively we define a least-squares estimate of sample partition for each of proteins, $\mathbf{c}_g^{*LS}$ by letting $\mathbf{c}_g^* = \mathbf{c}_{w_g}$ and using the posterior pairwise co-clustering probabilities of samples with respect to protein g . We use both summaries in the later results.

In addition to the point estimates of partitions, posterior inference under the NoB-LoC model provides a complete probabilistic description of uncertainties. We summarize uncertainty using a distance measure of two partitions:

$$H(\mathbf{w}, \mathbf{w}^{LS}) = \sum_{g=1}^G \sum_{g'=g}^G |d_{g,g'}^{\mathbf{w}} - d_{g,g'}^{\mathbf{w}^{LS}}|.$$

The posterior distribution of $H(\mathbf{w}, \mathbf{w}^{LS})$ summarizes the uncertainty in the clustering structure of proteins. Similarly, we can measure the uncertainty of the clustering structures in samples using $H(\mathbf{c}_s, \mathbf{c}_s^{LS})$, $s = 1, \dots, S^{LS}$ and $H(\mathbf{c}_g^*, \mathbf{c}_g^{*LS})$, $g = 1, \dots, G$.

3 Simulation

3.1 Validation

We examined the proposed model through a simulation study comparing model estimates with the simulation truth and with inference under three alternative methods, the plaid model (Lazzeroni and Owen, 2002; Turner *et al.*, 2005), sparse hierarchical clustering (Witten and Tibshirani, 2010), and the DCIM model (Freudenberg *et al.*, 2010). We simulated 20 proteins ($G = 20$) with 120 samples ($N = 120$). We assumed a simulation truth with three protein sets including an inactive protein set, ($s = 0, 1, 2$), where protein set 1 had 8 proteins and protein set 2 has 4 proteins. Protein set 1 partitioned the samples into four sample clusters including an inactive sample cluster ($k = 0, \dots, 3$). We generated sample cluster labels c_{1j} for protein set 1, assuming that a sample belonged to one of the four sample clusters with equal probability. Similarly, protein set 2 had three sample clusters including an inactive sample cluster ($k = 0, 1, 2$). Again, we generated c_{2j} assuming that a sample had equal probability of belonging to one of the three sample clusters. Following the definition of the inactive protein set, we did not generate sample clusters for the inactive protein set 0. Next, we generated θ_{kg}^* and y_{ig} given $\mathbf{w}$ and $\mathbf{c}_s$. The details of how θ_{kg}^* and y_{ig} are generated are described in the supplementary materials. One realization of θ_{kg}^* for protein sets 1 and 2 following the simulation setup is listed in Table 1. Finally, we generated $y_{ig} \sim N(\theta_{ig}, \sigma_g^2)$ where $\sigma_g = 0.3$.

Figure 3 shows the heatmaps of y_{ig} for each of the three true protein sets ($s = 0, 1, 2$). After rearranging proteins and rearranging samples within each protein set according to the simulation truth, we observed clear local clustering patterns in the data. For better presentation, in the figure, y_{ig} were rescaled to zero mean and unit variance in each column. In protein sets 1 and 2, the inactive samples are displayed in the first block of rows and show large variability in the color-coded expression levels. The active samples show more homogeneous colors (grey shades) within each sample cluster. In protein set 0, samples do not cluster and the corresponding protein expression levels show large variability.

Figure 4 shows the clustering results from hierarchical clustering. The global clustering of proteins (samples) is based on all samples (proteins). Therefore hierarchical clustering can not recover the simulation truth of the clustering.

Next, we implemented posterior inference under the proposed NoB-LoC model. We used the result from hierarchical clustering to initialize $\mathbf{w}$: We cut the dendrogram of the hierarchical clustering to have 12 protein clusters including 5 singleton clusters. For the initialization we combined the five singleton clusters to define an inactive protein set, $s = 0$. We fixed $\mu_{0g} = \mu_{1g} = \mu_{2g}$ at the sample median, $\text{med}(y_{ig}, i = 1, \dots, N)$, $\alpha_0 = 0.6$ and $\alpha_1 = 0.8$. We specified the hyperparameters a_{lg} , b_{lg} , a_g and b_g , by fixing the mean and variance of the inverse gamma priors for σ_g^2 and τ_{lg}^2 . Specifically, we matched $E(\tau_{0g}^2) = E(\tau_{1g}^2) = E(\tau_{2g}^2)$ with the sample variance of y_{ig} and set $\text{Var}(\tau_{0g}^2) = \text{Var}(\tau_{1g}^2) = \text{Var}(\tau_{2g}^2) = 0.4$. Also we centered $p(\sigma_g^2)$ by setting $E(\sigma_g^2)$ equal to the simulation truth and $\text{Var}(\sigma_g^2) = 0.4$. We then implemented posterior inference using MCMC posterior simulation. We ran the MCMC simulation over 20,000 iterations, discarding the first 5,000 iterations as burn-in.

The least-squares summary of the posterior on $\mathbf{w}$ was $\mathbf{w}^{LS} = (1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0)$. The estimated clustering $\mathbf{w}^{LS}$ grouped proteins 1–8 and 9–12 into two different protein sets, and the remaining proteins into the inactive protein set. Inference on the proteins sets perfectly recovered the simulation truth.

Conditional on $\mathbf{w}^{LS}$, we computed the least-squares estimates of sample clusters for the two protein sets, $\mathbf{c}_s^{LS}$, $s = 1, 2$ and compared the estimated cluster membership to the truth. Table 2 summarizes the results. The table reports the number of correct classifications and misclassifications for each sample cluster. Our inference identifies the true sample cluster membership under true protein sets 1 and 2 well. In particular, Table 2a shows six estimated sample clusters for protein set 1, with clusters (columns in Table 2a) $\{0, 1, 2, 3\}$ dominating and largely overlapping with the four true sample clusters of true protein set 1 (including the inactive one). Similar observations can be made for Table 2b. Figure 5 shows the heatmap of rearranged data according to the estimates $\mathbf{w}^{LS}$ and $\mathbf{c}_s^{LS}$. The display compares well with the simulation truth in Figure 3.

Figure 6a–c summarizes the uncertainties related to the point estimates of $\mathbf{w}$ and $\mathbf{c}_g^*$ for some selected proteins. We rescaled $H(\mathbf{w}, \mathbf{w}^{LS})$ and $H(\mathbf{c}_g^*, \mathbf{c}_g^{*LS})$ with the maximum possible distance from $\mathbf{w}^{LS}$ and $\mathbf{c}_g^{LS}$, that is, $G(G+1)/2$ and $N(N+1)/2$, respectively. Figure 6a shows that the posterior distribution is tightly supported around $\mathbf{w}^{LS}$ (indirectly also confirming the least squares estimate $\mathbf{w}^{LS}$ as a reasonable posterior summary). The larger variability in $\mathbf{c}_g^*$ for active proteins $g = 4$ and 10 , reflects the limited information for the nested clustering within each protein set. In contrast, inference on the protein clusters $\mathbf{w}$ reflects the combined information across all sample clusters $\mathbf{c}_s$. Also, the number of samples is much larger than the number of proteins, allowing for more uncertainty in the random partition. Figure 6d illustrates variabilities associated with $\mathbf{c}_g^{LS}$ for all the proteins.

Finally, we investigated sensitivity of the reported inference with respect to the hyperparameter choices, in particular the total mass parameters α and M that index the prior probability models (1) and (2) for the partition of proteins and samples, respectively. Details are reported in section C of the supplementary materials. In summary, we find little sensitivity with respect to α . The reported point estimate $\mathbf{w}^{LS}$ remains unchanged for $\{0.1, 1, 10\}$. In contrast, we find that $\mathbf{c}_s^{LS}$ changes with M . For large $M = 10$ posterior inference adds many small and singleton sample clusters. However, the additional clusters

correspond to truly inactive samples. The active sample clusters remain to be well identified. Alternative to a fixed choice of the total mass parameters, the model could be augmented with hyperpriors on M and if desired on $\cdot$.

3.2 Comparison with Alternative Methods

As comparison, we carried out inference for the same simulated data using three alternative methods, including the plaid model, sparse hierarchical clustering, and the DCIM model. The three methods are implemented in the *R* packages, *biclust*, *sparcl*, and *gimmR*, respectively. See the supplementary materials for figures with corresponding heatmaps. Here we only summarize the main features of inference under the three alternative approaches.

The plaid model identified four biclusters. Except for bicluster 1, the other three biclusters included active proteins and inactive proteins together or proteins in true protein set 1 and proteins in true protein set 2 together. Due to the failure to identify the true protein clusters, none of the identified biclusters matched with any true local clusters.

Sparse hierarchical clustering produced two sets of sample clusters. The first set was obtained through a direct implementation of the model in Witten and Tibshirani (2010). The second set of clusters was obtained after removing the effects found in the first clustering from the data, as suggested by the authors. Each of the partitions was based on its own chosen subset of proteins. The first clustering chose 10 proteins, $\{1, 3, 4, 5, 7, 8, 9, 10, 12, 14\}$, and the second clustering chose 8 proteins, $\{1, 5, 8, 13, 14, 15, 17, 20\}$. All the proteins in the true protein set 2 were chosen by the first clustering. We cut the resulting dendrogram to form three sample clusters, choosing the number of clusters to favor a match with the simulation truth that included $K = 3$ sample clusters under the true protein set 2. Similarly, the proteins chosen by the second clustering were in true protein set 1 and in the true inactive protein set. We cut the resulting dendrogram and formed four sample clusters. The sample cluster memberships of the first and second clusterings were compared to the true cluster memberships of protein sets 2 and 1, respectively. Table 3 summarizes the comparison results. The estimated sample partitions under sparse hierarchical clustering do not match well with the simulation truth, perhaps because sparse hierarchical clustering forces all samples to be assigned to a cluster. Sparse hierarchical clustering may also suffer from the inclusion of inactive proteins.

The DCIM model summarized posterior inference as two sets of global hierarchical clusters, one for proteins and one for samples. The DCIM model forms contexts of samples in which proteins are similarly clustered. The way how the DCIM model forms contexts is similar to the formation of protein sets in our model. We therefore transposed the data matrix before applying the DCIM model and reported the resulting clustering of proteins and the global partition of samples. We first cut the dendrogram for proteins to form three protein sets. We then considered the dendrogram for samples corresponding to the global sample clusters, separately for each of the protein sets to form protein-set-specific sample partitions. The three protein sets were identical to the protein sets characterized by w^{LS} under the NoB-LoC model. The global sample clustering under the DCIM model exhibited roughly five large sample clusters. Cutting the dendrogram for the global sample clustering into four sample clusters yielded a good sample partition for protein set 1, but not for protein set 2. Table 4 summarizes the estimated sample partitions for the first two protein sets. In Table 4a we see that the DCIM model recovers the simulation truth for the sample clustering under protein set 1, but Table 4b shows extensive mis-classification for the estimated sample clusters under protein set 2. Finally, we cut the dendrogram for protein set 3, the true inactive protein set, to form five sample clusters. This number was arbitrarily chosen after inspection of the dendrogram. The resulting sample clusters were noisy because protein set 3 was truly

inactive in the simulation truth (shown in Figure 3 in the supplementary materials). In contrast, the NoB-LoC model correctly identified this protein set as inactive and did not attempt to partition the samples.

Overall, inference under the NoB-LoC model compares favorably with the considered alternatives. Posterior inference did well in recovering the true clustering patterns.

3.3 Zero Enrichment

We consider one more alternative analysis of the simulated data to investigate the importance of explicitly modeling inactive proteins and samples. We replaced the zero-enriched Pólya urn in equations (1) and (2) with a regular Pólya urn (without zero-enrichment) and compared the simulation results under both setups. We used the same hyperparameters for the modified model. We also initialized $\mathbf{w}$ as before, except for combining 5 singleton clusters as one active protein set. We ran the MCMC simulation by iterating over all complete conditionals for 20,000 iterations, discarding the first 5,000 iterations as burn-in.

We computed the least-squares estimates $\mathbf{w}^{LS}$ of protein sets and the corresponding estimate $\mathbf{c}_s^{LS}$ of sample clusters. The modified model identified two protein sets shown in Figure 7. Protein set 1 included proteins 1–8 and 17 and protein set 2 included proteins 9–16 and 18–20.

Comparing the estimated partitions with the simulation truth, the model without zero-inflation failed to separate truly inactive proteins that did not induce any meaningful sample partition. In addition, the regular Pólya urn model generated many small and meaningless sample clusters under the two protein sets. Furthermore, inclusion of inactive proteins and samples inflated sampling variances, σ_g^2 . Posterior estimates of the sampling variance for the truly inactive proteins, $\hat{\sigma}_g^2$, $g = 13, \dots, 20$, were around 0.275, approximately three times larger than the simulation truth $\sigma_g^2 = 0.09$. The corresponding estimates under the NoB-LoC model, $\hat{\sigma}_g^2$ were around 0.079.

4 RPPA Data

To fit the NoB-LoC model for the RPPA data, we initialized $\mathbf{w}$ based on the results from hierarchical clustering. We cut the hierarchical clustering tree for proteins at a level that produced 13 protein clusters and combined all singleton protein clusters into an inactive protein set. We matched $\mu_{0g} = \mu_{1g} = \mu_{2g}$ with the sample median, $\text{med}(y_{ig}, i = 1, \dots, N)$ and fixed $\alpha_0 = 0.86$ and $\alpha_1 = 0.5$. To set the hyperparameters a_g , b_g , a_g and b_g , we chose values $E(\tau_{0g}^2) = E(\tau_{1g}^2) = E(\tau_{2g}^2)$ to match the sample variance of the y_{ig} and fixed $\text{Var}(\tau_{0g}^2) = \text{Var}(\tau_{1g}^2) = \text{Var}(\tau_{2g}^2) = 1$. We set $E(\sigma_g^2)$ equal to the sample variance of y_{ig} after trimming 5% from each tail and fixed $\text{Var}(\sigma_g^2) = 5$. We implemented inference using MCMC posterior simulation over 35,000 iterations, discarding the first 5,000 iterations as burn-in.

We computed the least-squares estimates of protein sets and the corresponding sample clusters, $\mathbf{w}^{LS}$ and $\mathbf{c}_s^{LS}$. The NoB-LoC model identified four active protein sets, each of which partitioned samples differently. The protein sets are shown in Figure 8. In the figure, each panel shows the composition of protein sets and the corresponding sample clustering, with the inactive sample clusters arranged in the first block of rows. The heatmaps of all the protein sets including the inactive protein set are shown in Figure 9 of the supplementary materials. The inferred sample clusters were clearly visible as homogeneous blocks of

protein expression. Figure 9 shows the posterior distributions of normalized $H(\mathbf{w}, \mathbf{w}^{LS})$ and $H(\mathbf{c}_g^*, \mathbf{c}_g^{*LS})$ for some selected proteins to summarize the uncertainties related to the point estimates of the protein and sample partitions. Panel (a) implies a large posterior probability for the reported point estimate $\mathbf{w}^{LS}$ (once again indirectly validating the least squares summary of the posterior distribution on $\mathbf{w}$). The other three panels, (b)–(d) summarize the uncertainties in the inference for $\mathbf{c}_g^*$, for arbitrarily selected proteins $g = 2, 19$, and 25 .

We compared the local sample clusters shown in Figure 8 with the traditionally used clinical subtypes determined by ER, PR and HER2. In the posterior estimated partition $\mathbf{w}^{LS}$, PR, ER and HER2 belonged to protein sets 1, 2 and 3, respectively. We found that under protein sets 1, 2, and 3 samples with subtype HR+ were roughly clustered together and not clustered with samples with subtypes HER2+ and TN. This agrees with the prior knowledge that subtype HR+ has very different characteristics from subtypes HER2+ and TN. However, overall there was little overlap between the local clusters identified by the NoB-LoC method and the clinical definition. As we already discussed earlier, in Section 1.2, this is no surprise, as the clinical definition of subtypes has not yielded clear and successful prognoses, suggesting poor understanding and perhaps inadequate knowledge of different subtypes of breast cancer. However, several interesting observations seem to suggest that our inference agrees with important knowledge and also implies new theory, which we discuss next.

The clustering results from the proposed model confirm some known breast cancer subtypes and genomic regulation mechanisms that have been recently reported by TCGA researchers (<http://cancergenome.nih.gov>) and others. First, protein set 3 contains proteins, PTEN, INPP4B, and HER2, for which inactivation (shown in red) has been observed among sample clusters. These sample clusters represent a breast cancer subtype, known as “basal-like”, characterized by the lack of the three proteins (The Cancer Genome Atlas Network, 2012). Both PTEN and INPP4B are phosphatases that can suppress the PI3K signaling pathway. The pathway is known to be involved in cell cycle progression and cell survival. Its excessive and aberrant activity is associated with tumorigenesis. Second, some proteins in the MAPK pathway including MEK1/2 and p38 are grouped with HER2 in protein set 3. It is consistent with the previous findings that in breast cancer cells HER2 activation results in activation of MAPK signaling pathway, eventually leading to cell proliferation (Oh *et al.*, 2001). Third, the inclusion of PR in protein set 1 with many phosphorylated PI3K pathway proteins (like AKT and mTOR), as shown in Figure 8, supports the regulation of PR expression through PI3K/AKT/mTOR signaling in breast cancer (Cui *et al.*, 2003). Fourth, the three biomarker proteins, ER, PR, and HER2, are assigned into three different protein sets, i.e. protein sets 2, 1, and 3, respectively. This suggests that the data supports three parallel partitions of the samples, with respect to protein sets containing the three markers, rather than one partition with respect to the three markers. This reflects the fact that all three, HER2, ER, and PR have their own signaling pathways, and these are signaling pathways for different genomic functions. This could explain why they define different partitions into phenotypes on patients.

In addition to these observations that are validated by recent breast cancer research, the reported results from local clustering also suggest some novel hypotheses for future studies. For example, protein set 2, as shown in Figure 8, includes both PI3K pathway proteins (e.g. 4EBP1 and PDK1) and MAPK pathway proteins (e.g. MAPK and S6) together. The co-expression of these proteins indicates potential co-regulation or crosstalk between the two signaling pathways involved in breast cancer development that requires further investigation. The classification of breast cancer subtypes is an important and ongoing topic (Curtis *et al.*, 2012; The Cancer Genome Atlas Network, 2012). The proposed inference on

local (nested) clustering has the potential to help biomedical researchers identify new subtypes based on molecular patterns.

For comparison we also considered inference under the the plaid model, sparse hierarchical clustering and the DCIM model for the same data set. The plaid model identified three biclusters. One cluster consisted of 12 proteins with 27 samples, the second had 12 proteins with 196 samples, and the third had three proteins with 70 samples. The samples in the three biclusters did not naturally correspond to any of the clinical subtypes.

We also applied the sparse hierarchical clustering approach. It chose all proteins for the first clustering and again all proteins except protein 18 for the second clustering. We cut the dendrograms to form three sample clusters, to favor a possible match with the three clinical subtypes. However, a comparison of the estimated sample partitions with the three cancer subtypes found no natural match.

Finally, we applied the DCIM model to the data. We cut the dendrogram of the protein clustering in the DCIM model to form four protein sets, and we cut the dendrogram of the global sample clustering in the DCIM model to form three sample clusters. Again, we chose the cut into three sample clusters to favor a possible match with the three clinical subtypes. However, the resulting sample partition did not show any natural match with the three clinical types. The first estimated protein set included all proteins in protein sets 1, 2 and 3 implied by w^{LS} under the NoB-LoC model, except for proteins 53 and 54. The NoB-LoC model assigned those proteins into three separate protein sets and allowed the three protein sets to partition samples differently. In contrast, the DCIM model inferred one common sample partition for those proteins.

We examined the results from the three alternative clustering methods (see Figures 9-11 in the supplementary materials) for support of the observations suggested by our inference described earlier and we found that they did not reflect the suggested observations. The plaid model failed to select almost half of the proteins including ER; the sparse hierarchical clustering produced a large cluster of proteins that included almost all biomarkers; and the DCIM method yielded noisy sample clusters for protein sets 2-4.

5 Conclusions

The proposed model extracts local clusters from a data matrix, by assuming that sample (row) partitions are nested with protein (column) sets. We achieve robust estimates of local clustering patterns by probabilistically excluding noisy proteins and samples from clustering. We use the zero-enriched Pólya urn as a prior model for a random partition of proteins and samples. If desired, any alternative random partition models could be substituted without substantially changing the approach. Alternative models could be used to represent different prior judgment on the number and relative sizes of the clusters. In addition, overlapping clusters could be introduced into the probability models, reflecting the notion that many proteins can feature in multiple biologic processes. The lack of such overlap is a limitation of the proposed model.

Another direction of possible extensions is the sampling model. The assumed normal sampling model could be replaced by other models without changes to the random partition models.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Yuan Ji and Peter Müller's research is partially supported by NIH R01 CA132897.

References

- Belacel N, Wang QC, Cuperlovic-Culf M. Clustering methods for microarray gene expression data. *OMICS: A Journal of Integrative Biology*. 2006; 10(4):507–531. [PubMed: 17233561]
- Blackwell D, MacQueen JB. Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*. 1973; 1:353–355.
- Carlin BP, Chib S. Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B: Methodological*. 1995; 57:473–484.
- Carvalho CM, Chang J, Lucas JE, Nevins JR, Wang Q, West M. High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics. *Journal of the American Statistical Association*. 2008; 103(484):1438–1456. [PubMed: 21218139]
- Cheng Y, Church GM. Biclustering of expression data. *Proceedings of the Eight International Conference on Intelligent Systems for Molecular Biology (ISBM)*. 2000; 8:93–103.
- Cui X, Zhang P, Deng W, Oesterreich S, Lu Y, Mills G, Lee A. Insulin-like growth factor-i inhibits progesterone receptor expression in breast cancer cells via the phosphatidylinositol 3-kinase/akt/mammalian target of rapamycin pathway: progesterone receptor as a potential indicator of growth factor activity in breast cancer. *Molecular Endocrinology*. 2003; 17(4):575–588. [PubMed: 12554765]
- Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, Gr'af S, Ha G, Haffari G, Bashashati A, Russell R, McKinney S, Group M, Langerd A, Green A, Provenzano E, Wishart G, Pinder S, Watson P, Markowitz F, Murphy L, Ellis I, Purushotham A, Brresen-Dale AL, Brenton JD, Tavaré S, Caldas C, Aparicio S. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*. 2012; 486:346–352. [PubMed: 22522925]
- Dahl, DB. Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model. In: Vannucci, M.; Do, KA.; Müller, P., editors. *Bayesian Inference for Gene Expression and Proteomics*. Cambridge University Press; 2006.
- Dunson DB. Nonparametric Bayes local partition models for random effects. *Biometrika*. 2009; 96(2):249–262. [PubMed: 23710074]
- Dunson DB, Xue Y, Carin L. The Matrix Stick-Breaking Process: Flexible Bayes Meta-Analysis. *Journal of the American Statistical Association*. 2008; 103(481):317–327.
- Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*. 2002; 97(458):611–631.
- Freudenberg JM, Sivaganesan S, Wagner M, Medvedovic M. A semi-parametric bayesian model for unsupervised differential co-expression analysis. *BMC Bioinformatics*. 2010; 11:234. [PubMed: 20459663]
- Jiang D, Tang C, Zhang A. Clustering analysis for gene expression data: a survey. *IEEE Transactions on Knowledge and Data Engineering*. 2004; 16:1370–1386.
- Lazzeroni L, Owen A. Paid Models for Gene Expression Data. *Statistica Sinica*. 2002; 12:61–86.
- Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, Shyr Y, Pietsenpol JA. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *The Journal of Clinical Investigation*. 2011; 121(7):27502767.
- Medvedovic M, Sivaganesan S. Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics*. 2002; 18(9):1194–206. [PubMed: 12217911]
- Medvedovic M, Yeung K, Bumgarner R. Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics*. 2004; 22(8):1222–32. [PubMed: 14871871]
- Meeds, E.; Roweis, S. Nonparametric bayesian biclustering Tech rep. Department of Computer Science, University of Toronto; 2007.
- Müller P, Quintana FA, Rosner GL. A product partition model with regression on covariates. *Journal of Computational and Graphical Statistics*. 2011; 20(1):260–278. [PubMed: 21566678]

- Oh AS, Lorant LA, Holloway JN, Miller DL, Kern FG, El-Ashry D. Hyperactivation of mapk induces loss of er expression in breast cancer cells. *Molecular Endocrinology*. 2001; 15(8):1344–1359. [PubMed: 11463858]
- Petrone S, Guindani M, Gelfand AE. Hybrid Dirichlet mixture models for functional data. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*. 2009; 71(4):755–782.
- Quintana FA. A predictive view of Bayesian clustering. *Journal of Statistical Planning and Inference*. 2006; 136(8):2407–2429.
- Rodríguez A, Dunson DB, Gelfand AE. The Nested Dirichlet Process. *Journal of the American Statistical Association*. 2008; 103(483):1131–1144.
- Rodríguez, A.; Ghosh, K. Modeling relational data using nested infinite relational models Tech rep. Department of Applied Mathematics and Statistics, Univeristy of California; Santa Cruz: 2012.
- Roy DM, Teh YW. The mondrian process. *Advances in Neural Information Processing Systems*. 2009; 21
- Sivaganesan S, Laud PW, Müller P. A Bayesian subgroup analysis with a zero-enriched Polya Urn scheme. *Statistics in Medicine*. 2011; 4:312–23. [PubMed: 21225894]
- The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*. 2012; 490:61–70. [PubMed: 23000897]
- Tibes R, Qiu Y, Lu Y, Hennessy B, Andreeff M, Mills GB, Kornblau SM. Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. *Molecular Cancer Therapeutics*. 2006; 5:2512–2521. [PubMed: 17041095]
- Turner H, Bailey T, Krzanowski W. Improved biclustering of microarray data demonstrated through systematic performance tests. *Computational Statistics & Data Analysis*. 2005; 48:235–254.
- Wade S, Mongelluzzo S, Petrone S. An enriched conjugate prior for bayesian nonparametric inference. *Bayesian Analysis*. 2011; 6(3):359–385.
- Witten DM, Tibshirani R. A framework for feature selection in clustering. *Journal of the American Statistical Association*. 2010; 105(490):713–726. [PubMed: 20811510]

Figure 1.
A heatmap of the RPPA data set. Dendrograms are estimated using hierarchical clustering.

Figure 2.

An illustration of the NoB-LoC model. The partition of proteins (columns) is highlighted by the protein sets, and the nested partition of samples (rows) is represented by sample clusters. Protein set 0 and sample cluster 0 are special inactive clusters.

Figure 3.

The heatmap of the three protein sets under the simulation truth. The samples are rearranged according to the simulation truth in protein sets 1 and 2. Protein set 0 is inactive. White horizontal lines divide the true sample clusters. The first block of top rows for each protein set consists of the inactive samples for that protein set.

Figure 4. Inference for the simulated data based on hierarchical clustering. The resulting protein and sample clusters are very different from the simulation truth.

Figure 5.

The heatmap of three estimated protein sets under the NoB-LoC model. The samples are rearranged according to the estimated cluster membership for protein sets 1 and 2. White horizontal lines divide the sample clusters. The first block of top rows for each active protein set consists of the inactive samples for that protein set.

Figure 6.

Inference for the simulated data based on NoB-LoC. Posterior distribution of scaled $H(\mathbf{w}, \mathbf{w}^{LS})$ and $H(\mathbf{c}_g^*, \mathbf{c}_g^{*LS})$ for (arbitrarily selected) proteins, $g = 4, 10$ in panels (a)-(c). Panel (d) is a boxplot of scaled $H(\mathbf{c}^*, \mathbf{c}^{*LS})$.

Figure 7.

The heatmap of two identified protein sets for the simulated data using the NoB-LoC model *without* zero-inflated Pólya urn. The samples are rearranged according to the estimated cluster membership for protein sets 1 and 2. White horizontal lines divide the sample clusters. The first block of top rows for each active protein set consists of the inactive samples for that protein set.

Figure 8.

Inference for the RPPA data based on NoB-LoC. Heatmaps of four identified protein sets (ps) by the NoB-LoC model. Notation “ps” refers to “protein set”. The samples are rearranged according to the estimated cluster memberships within each of active protein sets 1, 2, 3 and 4. White horizontal lines show the sample clusters. The first block of top rows for each active protein set are the inactive samples for that protein set. The protein names in red are cited in the text to relate the clustering results to new clinical findings.

Figure 9.

Inference for the RPPA data based on the NoB-LoC model. Posterior distribution of rescaled $H(\mathbf{w}, \mathbf{w}^{LS})$ and $H(\mathbf{c}_g^*, \mathbf{c}_g^{*LS})$ for some selected proteins, $g = 2, 19, 25$.

Table 1

True locations of the sample clusters for the simulated data.

(a) Locations for protein set 1											
	protein 1	protein 2	protein 3	protein 4	protein 5	protein 6	protein 7	protein 8			
cluster 1	-0.87	0.73	-0.26	-0.66	0.41	-0.37	-0.47	-0.01			
cluster 2	-0.50	-0.35	-0.73	0.48	-0.32	0.40	-0.99	0.52			
cluster 3	0.29	0.36	0.47	-0.52	-0.88	-0.05	0.80	0.96			

(b) Locations for protein set 2			
	protein 9	protein 10	protein 11
cluster 1	0.43	-0.93	-0.49
cluster 2	-0.84	0.58	0.29

	protein 12
cluster 1	0.44
cluster 2	-0.61

Inference for the simulated data based on NoB-LoC. Estimated sample cluster memberships c_1^{LS} and c_2^{LS} for protein sets 1 and 2 are compared with the simulation truth. The table reports the numbers of agreements and disagreements for sample clustering.

Table 2

(a) Protein set 1 of NoB-LoC							(b) Protein set 2 of NoB-LoC							
c_1^{LS}							c_2^{LS}							
c^{TRUE}	0	1	2	3	4	5	c^{TRUE}	0	1	2	3	4	5	6
0	28	0	1	1	2	1	0	26	4	1	2	1	1	1
1	0	26	0	0	0	0	1	0	38	0	0	0	0	0
2	0	0	28	0	0	0	2	0	0	45	0	0	0	0
3	0	0	0	33	0	0								

Table 3

Inference for the simulated data based on sparse hierarchical clustering. The table reports the numbers of agreements and disagreements between sparse hierarchical clustering and the simulation truth for sample clustering. The clusters under sparse hierarchical clustering are obtained by cutting the dendrograms.

(a) First clustering		(b) Second clustering						
c^{TRUE}	Estimated clusters			c^{TRUE}	Estimated clusters			
	1	2	3		1	2	3	4
0	22	5	9	0	16	8	9	0
1	17	22	0	1	2	17	0	7
2	27	0	18	2	1	11	16	0
				3	0	1	32	0

Table 4

Inference for the simulated data based on DCIM. The table reports the numbers of agreements and disagreements between DCIM and the simulation truth for sample clustering. The clusters under DCIM are obtained by cutting the dendrograms.

(a) Protein set 1 of DCIM					(b) Protein set 2 of DCIM				
Estimated clusters					Estimated clusters				
c^{TRUE}	1	2	3	4	c^{TRUE}	1	2	3	
$k=0$	31	0	1	1	$k=0$	16	10	10	
$k=1$	0	26	0	0	$k=1$	25	8	6	
$k=2$	0	0	28	0	$k=2$	19	16	10	
$k=3$	0	0	0	33					