BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 20 2013, pages 2610–2616
doi:10.1093/bioinformatics/btt425

Systems biology

Advance Access publication August 28, 2013

Bayesian consensus clustering

Eric F. Lock[1,2,*] and David B. Dunson[1]

[1]Department of Statistical Science, Duke University, Durham, NC 27708, USA and [2]Center for Human Genetics, Duke University Medical Center, Durham, NC 27710, USA

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: In biomedical research a growing number of platforms and technologies are used to measure diverse but related information, and the task of clustering a set of objects based on multiple sources of data arises in several applications. Most current approaches to multisource clustering either independently determine a separate clustering for each data source or determine a single 'joint' clustering for all data sources. There is a need for more flexible approaches that simultaneously model the dependence and the heterogeneity of the data sources.

Results: We propose an integrative statistical model that permits a separate clustering of the objects for each data source. These separate clusterings adhere loosely to an overall consensus clustering, and hence they are not independent. We describe a computationally scalable Bayesian framework for simultaneous estimation of both the consensus clustering and the source-specific clusterings. We demonstrate that this flexible approach is more robust than joint clustering of all data sources, and is more powerful than clustering each data source independently. We present an application to subtype identification of breast cancer tumor samples using publicly available data from The Cancer Genome Atlas.

Availability: R code with instructions and examples is available at http://people.duke.edu/%7Eel113/software.html.

Contact: Eric.Lock@duke.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on April 3, 2013; revised on June 18, 2013; accepted on July 18, 2013

1 INTRODUCTION

1.1 Motivation

Several fields of research now analyze *multisource* data (also called *multimodal* data), in which multiple heterogeneous datasets describe a common set of objects. Each dataset represents a distinct mode of measurement or domain.

While the methodology described in this article is broadly applicable, our primary motivation is the integrated analysis of heterogeneous biomedical data. The diversity of platforms and technologies that are used to collect genomic data, in particular, is expanding rapidly. Often multiple types of genomic data, measuring various biological components, are collected for a common set of samples. For example, The Cancer Genome Atlas (TCGA) is a large-scale collaborative effort to collect and catalog data from several genomic technologies. The integrative analysis of data from these disparate sources provides a more comprehensive understanding of cancer genetics and molecular biology.

Separate analyses of each data source may lack power and will not capture intersource associations. At the other extreme, a joint analysis that ignores the heterogeneity of the data may not capture important features that are specific to each data source. Exploratory methods that simultaneously model shared features and features that are specific to each data source have recently been developed as flexible alternatives (Lock *et al*., 2013; Löfstedt and Trygg, 2011; Ray *et al*., 2012; Zhou *et al*., 2012). The demand for such integrative methods motivates a dynamic area of statistics and bioinformatics.

This article concerns integrative clustering. Clustering is a widely used exploratory tool to identify similar groups of objects (for example, clinically relevant disease subtypes). Hundreds of general algorithms to perform clustering have been proposed. However, our work is motivated by the need for an integrative clustering method that is computationally scalable and robust to the unique features of each data source.

In Section 3.3, we apply our integrative clustering method to mRNA expression, DNA methylation, microRNA expression and proteomic data from TCGA for a common set of breast cancer tumor samples. These four data sources represent different but highly related and dependent biological components. Moreover, breast cancer tumors are recognized to have important distinctions that are present across several diverse genomic and molecular variables. A fully integrative clustering approach is necessary to effectively combine the discriminatory power of each data source.

1.2 Related work

Most applications of clustering multisource data follow one of two general approaches:

(1) Clustering of each data source separately, potentially followed by a post hoc integration of these separate clusterings.

(2) Combining all data sources to determine a single 'joint' clustering.

Under approach (1), the level of agreement between the separate clusterings may be measured by the *adjusted Rand index* (Hubert and Arabie, 1985) or a similar statistic. Furthermore, *consensus clustering* (also called *ensemble clustering*) can be used to determine an overall partition of the objects that agrees the most with the source-specific clusterings. Several objective

*To whom correspondence should be addressed

functions and algorithms to perform consensus clustering have been proposed [for a survey see Nguyen and Caruana (2007)]. Most of these methods do not inherently model uncertainty, and statistical models assume that the separate clusterings are known in advance (Wang *et al*., 2010, 2011). Consensus clustering is most commonly used to combine multiple clustering algorithms, or multiple realizations of the same clustering algorithm, on a single dataset. Consensus clustering has also been used to integrate multisource biomedical data (Cancer Genome Atlas Network, 2012). Such an approach is attractive in that it models source-specific features, yet still determines an overall clustering, which is often of practical interest. However, the two stage process of performing entirely separate clusterings followed by post hoc integration limits the power to identify and exploit shared structure (see Section 3.2 for an illustration of this phenomenon).

Approach (2) effectively exploits shared structure, at the expense of failing to recognize features that are specific to each data source. Within a model-based statistical framework, one can find the clustering that maximizes a joint likelihood. Assuming that each source is conditionally independent given the clustering, the joint likelihood is the product of the likelihood functions for each data source. This approach is used by Kormaksson *et al*. (2012) in the context of integrating gene expression and DNA methylation data. The *iCluster* method (Mo *et al*., 2013; Shen *et al*., 2009) performs clustering by first fitting a Gaussian latent factor model to the joint likelihood; clusters are then determined by K-means clustering of the factor scores. Rey and Roth (2012) propose a dependency-seeking model in which the goal is to find a clustering that accounts for associations across the data sources.

More flexible methods allow for separate but dependent source clusterings. Dependent models have been used to simultaneously cluster gene expression and proteomic data (Rogers *et al*., 2008), gene expression and transcription factor binding data (Savage *et al*., 2010) and gene expression and copy number data (Yuan *et al*., 2011). Kirk *et al*. (2012) describe a more general dependence model for two or more data sources. Their approach, called *Multiple Dataset Integration* (MDI), uses a statistical framework to cluster each data source while simultaneously modeling the pairwise dependence between clusterings. Savage *et al*. (2013) use MDI to integrate gene expression, methylation, microRNA and copy number data for glioblastoma tumor samples from TCGA. The pairwise dependence model does not explicitly model adherence to an overall clustering, which is often of practical interest.

2 METHODS

2.1 Finite Dirichlet mixture models

Here we briefly describe the finite Dirichlet mixture model for clustering a single dataset, with the purpose of laying the groundwork for the integrative model given in Section 2.2. Given data X_{n} for N objects ($n = 1, \ldots, N$), the goal is to partition these objects into at most K clusters. Typically X_n is a multidimensional vector, but we present the model in sufficient generality to allow for more complex data structures. Let $f(X_n|\theta)$ define a probability model for X_n given parameter(s) θ. For example, f may be a Gaussian density defined by the mean and variance $\theta = (\mu, \sigma^2)$. Each X_n is drawn independently from a mixture distribution with K components, specified by the parameters $\theta_1, \ldots, \theta_K$. Let $C_n \in \{1, \ldots, K\}$ represent the component corresponding to X_n, and π_k be the probability that an arbitrary object belongs to cluster k:

$$\pi_k = P(C_n = k).$$

Then, the generative model is

$$X_n \sim f(\cdot|\theta_k) \text{ with probability } \pi_k.$$

Under a Bayesian framework, one can put a prior distribution on $\Pi = (\pi_1, \ldots, \pi_K)$ and the parameter set $\Theta = (\theta_1, \ldots, \theta_K)$. It is natural to use a Dirichlet prior distribution for Π. Standard computational methods such as Gibbs sampling can then be used to approximate the posterior distribution for Π, Θ and $\mathbb{C} = (C_1, \ldots, C_N)$. The Dirichlet prior is characterized by a K-dimensional concentration parameter β of positive reals. Low prior concentration (for example, $\beta_k \leq 1$) will allow some of the estimated π_k to be small, and therefore N objects may not represent all K clusters. Letting $K \to \infty$ gives a *Dirichlet process*.

2.2 Integrative model

We extend the Dirichlet mixture model to accommodate data from M sources $\mathbb{X}_1, \ldots, \mathbb{X}_M$. Each data source is available for a common set of N objects, where X_{mn} represents data m for object n. Each data source requires a probability model $f_m(X_n|\theta_m)$ parametrized by θ_m. Under the general framework presented here, each $\mathbb{X}_m$ may have disparate structure. For example, X_{1n} may give an image where f_1 defines the spectral density for a Gaussian random field, while X_{2n} may give a categorical vector where f_2 defines a multivariate probability mass function.

We assume there is a separate clustering of the objects for each data source, but that these adhere loosely to an overall clustering. Formally, each X_{mn} $n = 1, \ldots, N$ is drawn independently from a K-component mixture distribution specified by the parameters $\theta_{m1}, \ldots, \theta_{mK}$. Let $L_{mn} \in \{1, \ldots, K\}$ represent the component corresponding to X_{mn}. Furthermore, let $C_n \in \{1, \ldots, K\}$ represent the overall mixture component for object n. The source-specific clusterings $\mathbb{L}_m = (L_{m1}, \ldots, L_{mN})$ are dependent on the overall clustering $\mathbb{C} = (C_1, \ldots, C_N)$:

$$P(L_{mn} = k|C_n) = \nu(k, C_n, \alpha_m)$$

where α_m adjusts the dependence function ν. The data $\mathbb{X}_m$ are independent of $\mathbb{C}$ conditional on the source-specific clustering $\mathbb{L}_m$. Hence, $\mathbb{C}$ serves only to unify $\mathbb{L}_1, \ldots, \mathbb{L}_M$. The conditional model is

$$P(L_{mn} = k|X_{mn}, C_n, \theta_{mk}) \propto \nu(k, C_n, \alpha_m) f_m(X_{mn}|\theta_{mk}).$$

Throughout this article, we assume ν has the simple form

$$\nu(L_{mn}, C_n, \alpha_m) = \begin{cases} \alpha_m \text{if} C_n = L_{mn} \\ \frac{1-\alpha_m}{K-1} \text{otherwise} \end{cases} \quad (1)$$

where $\alpha_m \in [\frac{1}{K}, 1]$ controls the adherence of data source m to the overall clustering. More simply α_m is the probability that $L_{mn} = C_n$. So, if $\alpha_m = 1$, then $\mathbb{L}_m = \mathbb{C}$. The α_m are estimated from the data together with $\mathbb{C}$ and $\mathbb{L}_1, \ldots, \mathbb{L}_m$. In practice we estimate each α_m separately, or assume that $\alpha_1 = \ldots = \alpha_M$ and hence each data source adheres equally to the overall clustering. The latter is favored when $M = 2$ for identifiability reasons. More complex models that permit dependence of the α_ms are also potentially useful.

Let π_k be the probability that an object belongs to the overall cluster k:

$$\pi_k = P(C_n = k).$$

We assume a Dirichlet(β) prior distribution for $\Pi = (\pi_1, \ldots, \pi_K)$. The probability that an object belongs to a given source-specific cluster follows directly:

$$P(L_{mn} = k|\Pi) = \pi_k \alpha_m + (1 - \pi_k)\frac{1 - \alpha_m}{K - 1}. \quad (2)$$

Table 1. Notation

N	Number of objects
M	Number of data sources
K	Number of clusters
$\mathbb{X}_m$	Data source m
X_{mn}	Data for object n, source m
f_m	Probability model for source m
θ_{mk}	Parameters for f_m, cluster k
p_m	Prior distribution for θ_{mk}
C_n	Overall cluster for object n
π_k	Probability that $C_n = k$
L_{mn}	Cluster specific to X_{mn}
ν	Dependence function for C_n and L_{mn}
α_m	Probability that $L_{mn} = C_n$

Moreover, a simple application of Bayes rule gives the conditional distribution of $\mathbb{C}$:

$$P(C_n = k|\mathbb{L}, \Pi, \alpha) \propto \pi_k \prod_{m=1}^{M} \nu(L_{mn}, k, \alpha_m),$$

where ν is defined as in (1).

The number of possible clusters K is the same for $\mathbb{L}_1, \ldots, \mathbb{L}_M$ and $\mathbb{C}$. The link function ν naturally aligns the cluster labels, as cases in which the clusterings are not well aligned (a permutation of the labels would give better agreement) will have low posterior probability. The number of clusters that are actually represented may vary, and generally the source-specific clusterings $\mathbb{L}_m$ will represent more clusters than $\mathbb{C}$, rather than vice versa. This follows from Equation (2) and is illustrated in Section 2 of the Supplementary Material. Intuitively if object n is not allocated to any overall cluster in data source m (i.e. $L_{mn} \notin \mathbb{C}$), then X_{mn} does not conform well to any overall pattern in the data.

Table 1 summarizes the mathematical notation used for the integrative model.

2.3 Marginal forms

Integrating over the overall clustering C gives the joint marginal distribution of $\mathbb{L}_1, \ldots, \mathbb{L}_M$:

$$P(\{L_{mn} = k_m\}_{m=1}^{M}|\Pi, \alpha) \propto \sum_{k=1}^{K} \pi_k \prod_{m=1}^{M} \nu(k_m, k, \alpha_m). \quad (3)$$

Under the assumption that $\alpha_1 = \ldots = \alpha_M$ the model simplifies:

$$P(\{L_{mn} = k_m\}_{m=1}^{M}|\Pi, \alpha) \propto \sum_{k=1}^{K} \pi_k U^{t_k} \quad (4)$$

where t_k is the number of clusters equal to k and $U = \frac{(K-1)\alpha_1}{1-\alpha_1} \geq 1$. This marginal form facilitates comparison with the MDI method for dependent clustering. In the MDI model $\phi_{ij} > 0$ control the strength of association between the clusterings $\mathbb{L}_i$ and $\mathbb{L}_j$:

$$P(\{L_{mn} = k_m\}_{m=1}^{M}|\tilde{\Pi}, \Phi) \propto \prod_{m=1}^{M} \tilde{\pi}_{mk_m} \prod_{\{i<j|k_i=k_j\}} (1 + \phi_{ij}) \quad (5)$$

where $\tilde{\pi}_{mk} = P(L_{mn} = k)$. For $K = 2$ and $\tilde{\pi}_{1\cdot} = \tilde{\pi}_{2\cdot}$, it is straightforward to show that (4) and (5) are functionally equivalent under a parameter substitution (see Section 3 of the Supplementary Material). There is no such general equivalence between the models for $K > 2$ or $M > 2$, regardless of restrictions on $\tilde{\Pi}$ and Φ. This is not surprising, as MDI gives a general model of pairwise dependence between clusterings rather than a model of adherence to an overall clustering.

2.4 Estimation

Here we present a general Bayesian framework for estimation of the integrative clustering model. We use a Gibbs sampling procedure to approximate the posterior distribution for the parameters introduced in Section 2.2. The algorithm is general in that we do not assume any specific form for the f_m and the parameters θ_{mk}. We use conjugate prior distributions for α_m, Π and (if possible) θ_{mk}.

- $\alpha_m \sim \text{TBeta}(a_m, b_m, \frac{1}{K})$, the $\text{Beta}(a_m, b_m)$ distribution truncated below by $\frac{1}{K}$. By default we choose $a_m = b_m = 1$, so that the prior for α_m is uniformly distributed between $\frac{1}{K}$ and 1.
- $\Pi \sim \text{Dirichlet}(\beta_0)$. By default we choose $\beta_0 = (1, 1, \ldots, 1)$, so that the prior for Π is uniformly distributed on the standard $(M-1)$-simplex.
- The θ_{mk} have prior distribution p_m. In practice, one should choose p_m so that sampling from the conditional posterior $p_m(\theta_{mk}|\mathbb{X}_m, \mathbb{L}_m)$ is feasible.

Markov chain Monte Carlo (MCMC) proceeds by iteratively sampling from the following conditional posterior distributions:

- $\Theta_m|\mathbb{X}_m, \mathbb{L}_m \sim p_m(\theta_{mk}|\mathbb{X}_m, \mathbb{L}_m)$ for $k = 1, \ldots, K$.
- $\mathbb{L}_m|\mathbb{X}_m, \Theta_m, \alpha_m, \mathbb{C} \sim P(k|X_{mn}, C_n, \theta_{mk}, \alpha_m)$ for $n = 1, \ldots, N$, where
$$P(k|X_{mn}, C_n, \Theta_m) \propto \nu(k, C_n, \alpha_m) f_m(X_{mn}|\theta_{mk}).$$
- $\alpha_m|\mathbb{C}, \mathbb{L}_m \sim \text{TBeta}(a_m + \tau_m, b_m + N - \tau_m, \frac{1}{K})$, where τ_m is the number of samples n satisfying $L_{mn} = C_n$.
- $\mathbb{C}|\mathbb{L}_m, \Pi, \alpha \sim P(k|\Pi, \{L_{mn}, \alpha_m\}_{m=1}^{M})$ for $n = 1, \ldots, N$, where
$$P(k|\Pi, \{L_{mn}, \alpha_m\}_{m=1}^{M}) \propto \pi_k \prod_{m=1}^{M} \nu(k, L_{mn}, \alpha_m)$$
- $\Pi|\mathbb{C} \sim \text{Dirichlet}(\beta_0 + \rho)$, where ρ_k is the number of samples allocated to cluster k in $\mathbb{C}$.

This algorithm can be suitably modified under the assumption that $\alpha_1 = \ldots = \alpha_M$ (see Section 1.2 of the Supplementary Material).

Each sampling iteration produces a different realization of the clusterings $\mathbb{C}, \mathbb{L}_1, \cdots, \mathbb{L}_m$, and together these samples approximate the posterior distribution for the overall and source-specific clusterings. However, a point estimate may be desired for each of $\mathbb{C}, \mathbb{L}_1, \cdots, \mathbb{L}_m$ to facilitate interpretation of the clusters. In this respect, methods that aggregate over the MCMC iterations to produce a single clustering, such as that described in Dahl (2006), can be used.

It is possible to derive a similar sampling procedure using only the marginal form for the source-specific clusterings given in Equation (3). However, the overall clustering C is also of interest in most applications. Furthermore, incorporating C into the algorithm can actually improve computational efficiency dramatically, especially if M is large. As presented, each MCMC iteration can be completed in $O(MNK)$ operations. If the full joint marginal distribution of $L_1, \ldots, L_M$ is used the computational burden increases exponentially with M (this presents a bottleneck for the MDI method).

For each iteration, C_n is determined randomly from a distribution that gives higher probability to clusters that are prevalent in $\{L_{1n}, \ldots, L_{mn}\}$. In this sense, $\mathbb{C}$ is determined by a random consensus clustering of the source-specific clusterings. Hence, we refer to this approach as *Bayesian consensus clustering* (BCC). BCC differs from traditional consensus clustering in three key aspects.

(1) Both the source-specific clusterings and the consensus clustering are modeled in a statistical way that allows for uncertainty in all parameters.

(2) The source-specific clusterings and the consensus clustering are estimated simultaneously, rather than in two stages. This permits

borrowing of information across sources for more accurate cluster assignments.

(3) The strength of association to the consensus clustering for each data source is learned from the data and accounted for in the model.

We have developed software for the R environment for statistical computing (R Development Core Team, 2012) to perform BCC on multivariate continuous data using a Normal-Gamma conjugate prior distribution for cluster-specific means and variances. Full computational details for this implementation are given in Section 1.1 of the Supplementary Material. This software is open source and may be modified for use with alternative likelihood models (e.g. for categorical or functional data).

2.5 Choice of K

One can infer the number of clusters in the model by specifying a large value for the maximum number of clusters K, for example $K = N$. The number of clusters realized in $\mathbb{C}$ and the $\mathbb{L}_m$ may still be small. However, we find that this is not the case for high-dimensional structured data such as that used for the genomics application in Section 3.3. The model tends to select a large number of clusters even if the Dirichlet prior concentration parameters β_0 are small. The number of clusters realized using a Dirichlet process increases with the sample size; hence, if the number of mixture component is indeed finite, the estimated number of clusters is inconsistent as $N \to \infty$ (Miller and Harrison, 2013). This is undesirable for exploratory applications in which the goal is to identify a small number of interpretable clusters.

Alternatively, we consider a heuristic approach that selects the value of K that gives maximum adherence to an overall clustering. For each K, the estimated adherence parameters $\alpha_m \in [\frac{1}{K}, 1]$ are mapped to the unit interval by the linear transformation

$$\alpha_m^* = \frac{K\alpha_m - 1}{K - 1}.$$

We then select the value of K that results in the highest mean adjusted adherence

$$\bar{\alpha}^* = \frac{1}{M}\sum_{m=1}^{M} \alpha_m^*.$$

This approach will generally select a small number of clusters that reveal shared structure across the data sources.

3 RESULTS

3.1 Accuracy of $\hat{\alpha}$

We find that with reasonable signal the α_m can generally be estimated with accuracy and without substantial bias. To illustrate, we generate simulated datasets $\mathbb{X}_1 : 1 \times 200$ and $\mathbb{X}_2 : 1 \times 200$ as follows:

(1) Let $\mathbb{C}$ define two clusters, where $C_n = 1$ for $n \in \{1, \ldots, 100\}$ and $C_n = 2$ for $n \in \{101, \ldots, 200\}$.

(2) Draw α from a Uniform(0.5,1) distribution.

(3) For $m = 1, 2$ and $n = 1, \ldots, 200$, generate $L_{mn} \in \{1, 2\}$ with probabilities $P(L_{mn} = C_n) = \alpha$ and $P(L_{mn} \neq C_n) = 1 - \alpha$.

(4) For $m = 1, 2$, draw values X_{mn} from a Normal(1.5,1) distribution if $L_{mn} = 1$ and from a Normal($-1.5, 1$) distribution if $L_{mn} = 2$.

We generate 100 realizations of the above simulation, and estimate the model via BCC for each realization. We assume $\alpha_1 = \alpha_2$ in our estimation and use a uniform prior; further computational details are given in Section 4 of the Supplementary Material. Figure 1 displays $\hat{\alpha}$, the best estimate for both α_1 and α_2, versus the true α for each realization. The point estimate displayed is the mean over MCMC draws, and we also display a 95% credible interval based on the 2.5–97.5 percentiles of the MCMC draws. The estimated $\hat{\alpha}$ are generally close to the true α, and the credible interval contains the true value in 91 of 100 simulations. See Section 4 of the Supplementary Material for a more detailed study, including a simulation illustrating the effect of the prior distribution on $\hat{\alpha}$.

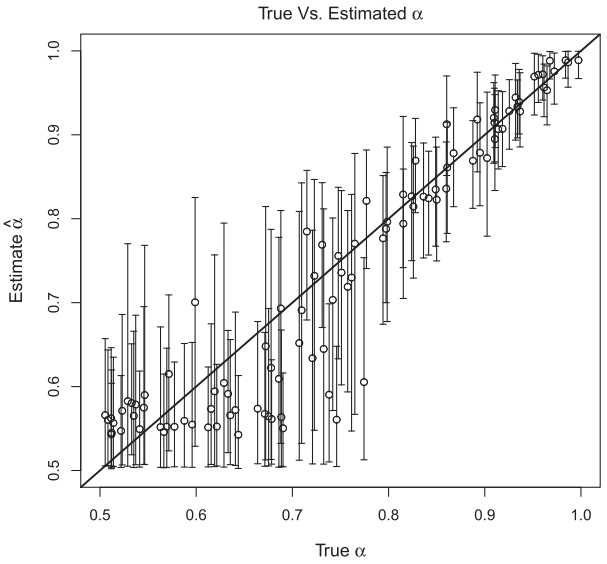

Fig. 1. Estimated $\hat{\alpha}$ versus true α for 100 randomly generated simulations. For each simulation, the mean value $\hat{\alpha}$ is shown with a 95% credible interval

3.2 Clustering accuracy

To illustrate the flexibility and advantages of BCC in terms of clustering accuracy, we generate simulated data sources $\mathbb{X}_1$ and $\mathbb{X}_2$ as in Section 3.1 but with Normal(1,1) and Normal(-1,1) as our mixture distributions. Hence, the signal distinguishing the two clusters is weak enough so that there is substantial overlap within each simulated data source. We generate 100 simulations and compare the results for four model-based clustering approaches:

(1) Separate clustering, in which a finite Dirichlet mixture model is used to determine a clustering separately for $\mathbb{X}_1$ and $\mathbb{X}_2$.

(2) Joint clustering, in which a finite Dirichlet mixture model is used to determine a single clustering for the concatenated data $[\mathbb{X}_1' \, \mathbb{X}_2']'$.

(3) Dependent clustering, in which we model the pairwise dependence between each data source, in the spirit of MDI.

(4) Bayesian consensus clustering.

The full implementation details for each method are given in Section 5 of the Supplementary Material.

We consider the relative error for each model in terms of the average number of incorrect cluster assignments:

$$\text{Source error} = \frac{\sum_{m=1}^{M}\sum_{n=1}^{N}\mathbb{1}\{\hat{L}_{mn} \neq L_{mn}\}}{MN},$$

$$\text{Overall error} = \frac{\sum_{n=1}^{N}\mathbb{1}\{\hat{C}_n \neq C_n\}}{N},$$

where $\mathbb{1}$ is the indicator function. For joint clustering, the source clusters $\widehat{\mathbb{L}}_m$ are identical. For separate and dependent clustering, we determine an overall clustering by maximizing the posterior expected adjusted Rand index (Fritsch and Ickstadt, 2009) of the source clusterings.

The relative error for each clustering method with $M=2$ and $M=3$ sources is shown in Figure 2. Smooth curves are fit to the results for each method using LOESS local regression (Cleveland, 1979) and display the relative clustering error for each method as a function of α. Not surprisingly, joint clustering performs well for $\alpha \approx 1$ (perfect agreement) and separate clustering performs well when $\alpha \approx 0.5$ (no relationship). BCC and dependent clustering learn the level of cluster agreement, and hence serve as a flexible bridge between these two extremes. Dependent clustering does not perform as well with $M=3$ sources, as the pairwise dependence model does not assume an overall clustering and therefore has less power to learn the underlying structure for $M>2$.

3.3 Application to genomic data

We apply BCC to multisource genomic data on breast cancer tumor samples from TCGA. For a common set of 348 tumor samples, our full dataset includes

- RNA gene expression (GE) data for 645 genes.
- DNA methylation (ME) data for 574 probes.
- miRNA expression (miRNA) data for 423 miRNAs.
- Reverse phase protein array (RPPA) data for 171 proteins.

These four data sources are measured on different platforms and represent different biological components. However, they all represent genomic data for the same sample set and it is reasonable to expect some shared structure. These data are publicly available from the TCGA Data Portal. See http://people.duke.edu/%7Eel113/software.html for R code to completely reproduce the following analysis, including instructions on how to download and process these data from the TCGA Data Portal.

Breast cancer is a heterogeneous disease and is therefore a natural candidate for clustering. Previous studies have found

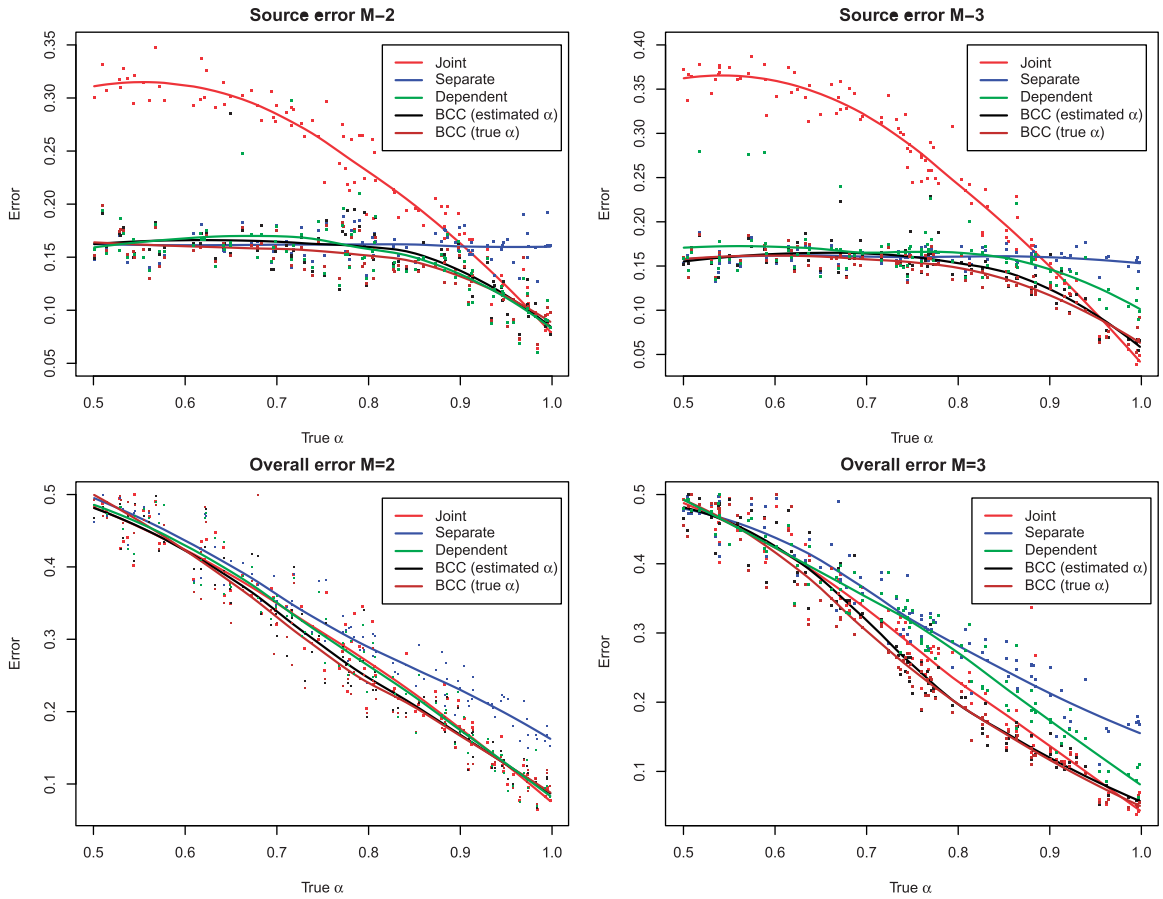

Fig. 2. Source-specific and overall clustering error for 100 simulations with $M=2$ and $M=3$ data sources, shown for joint clustering, separate clustering, dependent clustering, BCC and BCC using the true α. A LOESS curve displays clustering error as a function of α for each method

anywhere from 2 (Duan, 2013) to 10 (Curtis *et al*., 2012) distinct clusters based on a variety of characteristics. In particular, 4 comprehensive sample subtypes were previously identified based on a multisource consensus clustering of the TCGA data (Cancer Genome Atlas Network, 2012). These correspond closely to the well-known molecular subtypes Basal, Luminal A, Luminal B and HER2. These subtypes were shown to be clinically relevant, as they may be used for more targeted therapies and prognosis.

We use the heuristic described in Section 2.5 to select the number of clusters for BCC, with intent to determine a clustering that is well-represented across the four genomic data sources. We select $K = 3$ clusters, and posterior probability estimates were converted to hard clusterings via Dahl (2006) to facilitate comparison and visualization. Table 2 shows a matching matrix comparing the overall clustering $\mathbb{C}$ with the comprehensive subtypes defined by TCGA, as well as summary data for the BCC clusters.

The TCGA and BCC clusters show different structure but are not independent (P-value < 0.01; Fisher's exact test). BCC cluster 1 corresponds to the Basal subtype, which is characterized by basal-like expression and a relatively poor clinical prognosis. BCC cluster 2 is primarily a subset of the Luminal A samples, which are genomically and clinically heterogeneous. DNA copy number alterations, in particular, are a source of diversity for Luminal A. On independent datasets Curtis *et al*. (2012) and Jönsson *et al*. (2010) identify a subgroup of Luminal A that is characterized by fewer copy number alterations and a more favorable clinical prognosis (clusters *IntClust 3* and *Luminal-simple*, respectively). As a measure of copy number activity, we compute the fraction of the genome altered (FGA) as described in Cancer Genome Atlas Network (2012) Supplementary Section VII (with threshold $T = 0.50$) for each BCC cluster. Clusters 1 and 3 had an FGA above 0.2, while Cluster 2 had an FGA of 0.10 (Table 2). For comparison, those Luminal A samples that were not included in Cluster 2 had a substantially higher average FGA of 0.17 ± 0.02. Cluster 3 primarily includes those samples that are receptor (estrogen and/or progesterone) positive and have higher FGA. These results suggest that copy number variation may contribute to breast tumor heterogeneity across several genomic sources.

Table 2. BCC cluster versus TCGA comprehensive subtype matching matrix and summary data for BCC clusters

	BCC cluster		
	1	2	3
TCGA subtype			
1 (Her2)	13	6	20
2 (Basal)	66	2	4
3 (Lum A)	3	80	78
4 (Lum B)	0	3	73
5-year survival	0.67 ± 0.20	0.94 ± 0.08	0.81 ± 0.11
FGA	0.22 ± 0.04	0.10 ± 0.02	0.20 ± 0.02
ER+	13%	92%	94%
PR+	7%	86%	75%
HER2+	15%	12%	18%
8p11 amplification	32%	19%	42%
8q24 amplification	79%	39%	67%
5q13 deletion	61%	3%	14%
16q23 deletion	19%	66%	61%

Note: Summary data includes 5-year survival probabilities using the Kaplan–Meier estimator, with 95% confidence interval; mean fraction of the genome altered (FGA) using threshold $T = 0.5$, with 95% confidence interval; receptor status for estrogen (ER), progesteron (PR) and human epidermal growth factor 2 (HER2); and copy number status for amplification at sites 8p11 and 8q23 and deletion at sites 5q13 and 16q23.

Figure 3 provides a point-cloud view of each dataset given by a scatter plot of the first two principal components. The overall and source-specific cluster index is shown for each sample, as well as a point estimate and ~95% credible interval for the adherence parameter α. The GE data has by far the highest adherence to the overall clustering ($\alpha = 0.91$); this makes biological sense, as RNA expression is thought to have a direct causal relationship with each of the other three data sources. The four data sources show different sample structure, and the source-specific clusters are more well-distinguished than the overall clusters in each plot. However, the overall clusters are clearly represented to some degree in all four plots. Hence, the flexible, yet integrative, approach of BCC seems justified for these data.

Further details regarding the above analysis are given in Section 6 of the Supplementary Material. These include the prior specifications for the model, charts that illustrate mixing over the MCMC draws, a comparison of the source-specific clusterings L_{mn} to source-specific subtypes defined by TCGA, clustering heatmaps for each data source and short-term survival curves for each overall cluster.

4 DISCUSSION

This work was motivated by the perceived need for a general, flexible and computationally scalable approach to clustering multisource biomedical data. We propose BCC, which models both an overall clustering and a clustering specific to each data source. We view BCC as a form of consensus clustering, with advantages over traditional methods in terms of modeling uncertainty and the ability to borrow information across sources.

The BCC model assumes a simple and general dependence between data sources. When an overall clustering is not sought, or when such a clustering does not make sense as an assumption, a more general model of cluster dependence (such as MDI) may be more appropriate. Furthermore, a context-specific approach may be necessary when more is known about the underlying dependence of the data. For example, Nguyen and Gelfand (2011) exploit functional covariance models for time-course data to determine overall and time-specific clusters.

Our implementation of BCC assumes the data are normally distributed with cluster-specific mean and variance parameters. It is straightforward to extend this approach to more complex clustering models. In particular, models that assume clusters exist on a sparse feature set (Tadesse *et al*., 2005) or allow for more general covariance structure (Ghahramani and Beal, 1999) are growing in popularity.

While we focus on multisource biomedical data, the applications of BCC are potentially widespread. In addition to multisource data, BCC may be used to compare clusterings from different statistical models for a single homogeneous dataset.

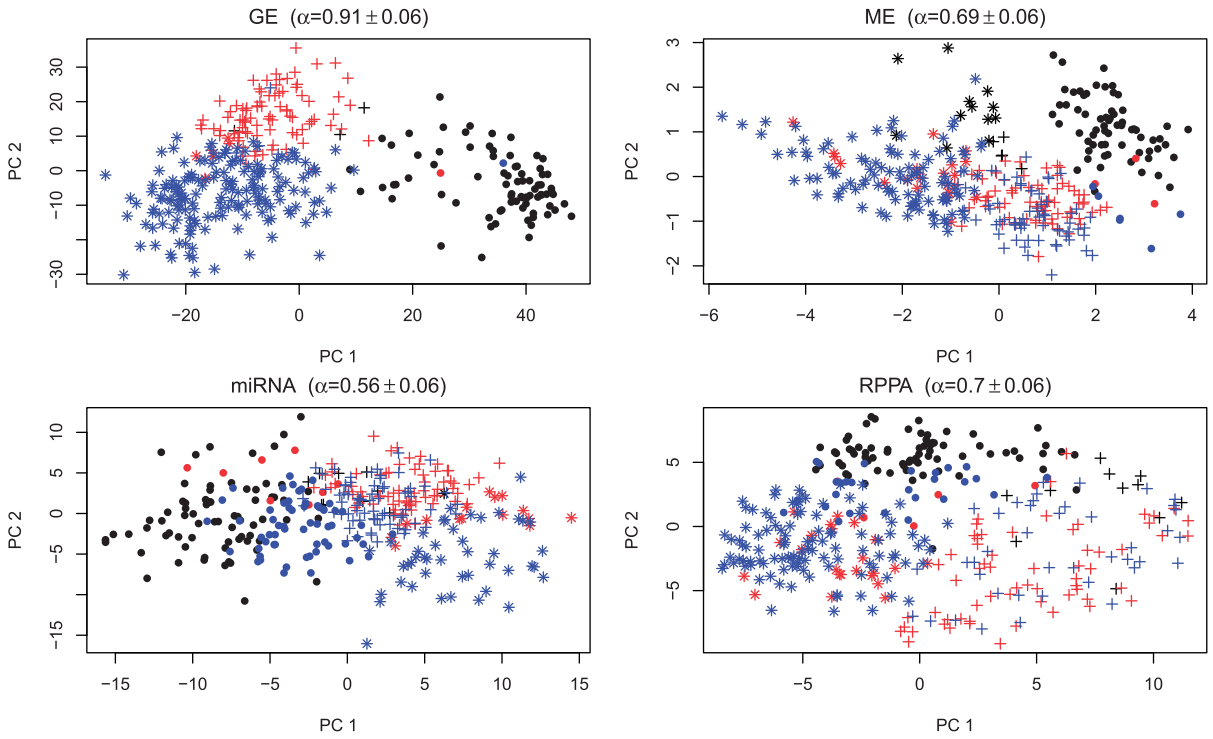

Fig. 3. PCA plots for each data source. Sample points are colored by overall cluster; cluster 1 is black, cluster 2 is red and cluster 3 is blue. Symbols indicate source-specific cluster; cluster 1 is indicated by filled circles, cluster 2 is indicated by plus signs and cluster 3 is indicated by asterisks

Funding: National Institute of Environmental Health Sciences (NIEHS) (R01-ES017436).

Conflict of Interest: none declared.

REFERENCES

Cancer Genome Atlas Network. (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.

Cleveland,W.S. (1979) Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.*, **74**, 829–836.

Curtis,C. *et al.* (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, **486**, 346–352.

Dahl,D. (2006) *Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model*. Cambridge University Press, Cambridge, UK.

Duan,Q. *et al.* (2013) Metasignatures identify two major subtypes of breast cancer. *CPT Pharmacom. Syst. Pharmacol.*, **3**, e35.

Fritsch,A. and Ickstadt,K. (2009) Improved criteria for clustering based on the posterior similarity matrix. *Bayesian Anal.*, **4**, 367–391.

Ghahramani,Z. and Beal,M.J. (1999) Variational inference for bayesian mixtures of factor analysers. In: Solla,S.A. *et al.* (ed.) *Advances in Neural Information Processing Systems 12, [NIPS Conference, Denver, Colorado, USA, November 29–December 4, 1999]*. The MIT Press, Cambridge, MA, USA, pp. 449–455.

Hubert,L. and Arabie,P. (1985) Comparing partitions. *J. Classif.*, **2**, 193–218.

Jönsson,G. *et al.* (2010) Genomic subtypes of breast cancer identified by array-comparative genomic hybridization display distinct molecular and clinical characteristics. *Breast Cancer Res.*, **12**, R42.

Kirk,P. *et al.* (2012) Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*, **28**, 3290–3297.

Kormaksson,M. *et al.* (2012) Integrative model-based clustering of microarray methylation and expression data. *Ann. Appl. Stat.*, **6**, 1327–1347.

Lock,E. *et al.* (2013) Joint and Individual Variation Explained (JIVE) for integrated analysis of multiple data types. *Ann. Appl. Stat.*, **7**, 523–542.

Löfstedt,T. and Trygg,J. (2011) Onplsa novel multiblock method for the modelling of predictive and orthogonal variation. *J. Chemom.*, **25**, 441–455.

Miller,J.W. and Harrison,M.T. (2013) A simple example of dirichlet process mixture inconsistency for the number of components. *arXiv preprint arXiv:1301.2708*.

Mo,Q. *et al.* (2013) Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc. Natl Acad. Sci. USA*, **110**, 4245–4250.

Nguyen,N. and Caruana,R. (2007) Consensus clusterings. In: *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM 2007), October 28-31, 2007, Omaha, Nebraska, USA*. pages 607–612. IEEE Computer Society.

Nguyen,X. and Gelfand,A.E. (2011) The Dirichlet labeling process for clustering functional data. *Stat. Sin.*, **21**, 1249–1289.

R Development Core Team. (2012) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ray,P. *et al.* (2012) Bayesian joint analysis of heterogeneous data. Preprint.

Rey,M. and Roth,V. (2012) Copula mixture model for dependency-seeking clustering. In: Langford,J. and Pineau,J. (eds) *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*. ICML'12, p. 927–934, New York, NY, Omnipress.

Rogers,S. *et al.* (2008) Investigating the correspondence between transcriptomic and proteomic expression profiles using coupled cluster models. *Bioinformatics*, **24**, 2894–2900.

Savage,R.S. *et al.* (2010) Discovering transcriptional modules by bayesian data integration. *Bioinformatics*, **26**, i158–i167.

Savage,R.S. *et al.* (2013) Identifying cancer subtypes in glioblastoma by combining genomic, transcriptomic and epigenomic data. *arXiv preprint arXiv:1304.3577*.

Shen,R. *et al.* (2009) Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, **25**, 2906–2912.

Tadesse,M.G. *et al.* (2005) Bayesian variable selection in clustering high-dimensional data. *J. Am. Stat. Assoc.*, **100**, 602–617.

Wang,H. *et al.* (2011) Bayesian cluster ensembles. *Stat. Anal. Data Mining*, **4**, 54–70.

Wang,P. *et al.* (2010) Nonparametric bayesian clustering ensembles. In: *Machine Learning and Knowledge Discovery in Databases*. Springer, Berlin - Heidelberg, pp. 435–450.

Yuan,Y. *et al.* (2011) Patient-specific data fusion defines prognostic cancer subtypes. *PLoS Comput. Biol.*, **7**, e1002227.

Zhou,G. *et al.* (2012) Common and individual features analysis: beyond canonical correlation analysis. *Arxiv preprint arXiv:1212.3913*.