Boluki *et al. BMC Bioinformatics* 2019, **20**(Suppl 12):321
https://doi.org/10.1186/s12859-019-2832-3

BMC Bioinformatics

RESEARCH Open Access

Optimal clustering with missing values

Shahin Boluki[1†], Siamak Zamani Dadaneh[1†], Xiaoning Qian[1,2] and Edward R. Dougherty[1,2*]

From International Workshop on Computational Network Biology: Modeling, Analysis and Control
Washington, D.C., USA. 29 August 2018

Abstract

Background: Missing values frequently arise in modern biomedical studies due to various reasons, including missing tests or complex profiling technologies for different omics measurements. Missing values can complicate the application of clustering algorithms, whose goals are to group points based on some similarity criterion. A common practice for dealing with missing values in the context of clustering is to first impute the missing values, and then apply the clustering algorithm on the completed data.

Results: We consider missing values in the context of optimal clustering, which finds an optimal clustering operator with reference to an underlying random labeled point process (RLPP). We show how the missing-value problem fits neatly into the overall framework of optimal clustering by incorporating the missing value mechanism into the random labeled point process and then marginalizing out the missing-value process. In particular, we demonstrate the proposed framework for the Gaussian model with arbitrary covariance structures. Comprehensive experimental studies on both synthetic and real-world RNA-seq data show the superior performance of the proposed optimal clustering with missing values when compared to various clustering approaches.

Conclusion: Optimal clustering with missing values obviates the need for imputation-based pre-processing of the data, while at the same time possessing smaller clustering errors.

Keywords: Clustering, Missing data, Optimal design, Pattern recognition

Background

Clustering has been a mainstay of genomics since the early days of gene-expression microarrays [1]. For instance, expression profiles can be taken over various tissue samples and then clustered according to the expression levels for each sample, the aim being to discriminate pathologies based on their differential patterns of gene expression [2]. In particular, model-based clustering, which assumes that the data are generated by a finite mixture of underlying probability distributions, has gained popularity over heuristic clustering algorithms, for which there is no concrete way of determining the number of clusters or the best clustering method [3]. Model-based clustering methods [4] provide more robust criteria for selecting the appropriate number of clusters. For example, in a Bayesian framework, utilizing Bayes Factor can incorporate both *a priori* knowledge of different models, and goodness of fit of the parametric model to the observed data. Moreover, nonparametric models such as Dirichlet-process mixture models [5] provide a more flexible approach for clustering, by automatically learning the number of components. In small-sample settings, model-based approaches that incorporate model uncertainty have proved successful in designing robust operators [6–9], and in objective-based experiment design to expedite the discovery of such operators [10–12].

Whereas classification theory is grounded in feature-label distributions with the error being the probability that the classifier mislabels a point [6, 13]; clustering algorithms operate on random labeled point processes (RLPPs) with error being the probability that a point will be placed into the wrong cluster (partition) [14]. An optimal (Bayes) clusterer minimizes the clustering error and

*Correspondence: edward@ece.tamu.edu
†Shahin Boluki and Siamak Zamani Dadaneh contributed equally to this work.
[1]Department of Electrical and Computer Engineering, Texas A&M University, MS3128 TAMU, 77843 College Station, TX, USA
[2]TEES-AgriLife Center for Bioinformatics & Genomic Systems Engineering, 77843 College Station, TX, USA

can be found with respect to an appropriate representation of the cluster error [15].

A common problem in clustering is the existence of missing values. These are ubiquitous with high-throughput sequencing technologies, such as microarrays [16] and RNA sequencing (RNA-seq) [17]. For instance, with microarrays, missing data can occur due to poor resolution, image corruption, or dust or scratches on the slide [18], while for RNA-seq, the sequencing machine may fail to detect genes with low expression levels owing to the random sampling nature of sequencing technologies. As a result of these missing data mechanisms, gene expression data from microarray or RNA-seq experiments are usually in the form of large matrices, with rows and columns corresponding to genes and experimental conditions or different subjects, respectively, with some values missing. Imputation methods, such as *MICE* [19], *Amelia II* [20] and *missForest* [21], are usually employed to complete the data matrix before clustering analysis; however, in small-sample settings, which are common in genomic applications, these methods face difficulties, including colinearity due to potential high correlation between genes in samples, which precludes the successful imputation of missing values.

In this paper we follow a different direction by incorporating the generation of missing values with the original generating random labeled point process, thereby producing a new RLPP that generates the actual observed points with missing values. The optimal clusterer in the context of missing values is obtained by marginalizing out the missing features in the new RLPP. One potential challenge arising here is that in the case of missing values with general patterns, conducting the marginalization can be computationally intractable, and hence resorting to approximation methods such as Monte Carlo integration is necessary.

Although the proposed framework for optimal clustering can incorporate the probabilistic modeling of arbitrary types of missing data mechanisms, to facilitate analysis, throughout this work we assume data are missing completely at random (MCAR) [22]. In this scenario, the parameters of the missingness mechanism are independent of other model parameters and therefore vanish after the expectation operation in the calculation of the posterior of label functions for clustering assignment.

We derive the optimal clusterer for different scenarios in which features are distributed according to multivariate Gaussian distributions. The performance of this clusterer is compared to various methods, including *k*-POD [23] and fuzzy *c*-means with optimal completion strategy [24], which are methods for directly clustering data with missing values, and also *k*-means [25], fuzzy *c*-means [26] and hierarchical clustering [27] with the missing values imputed. Comprehensive simulations based on synthetic data show the superior performance of the proposed framework for clustering with missing values over a range of simulation setups. Moreover, evaluations based on RNA-seq data further verify the superior performance of the proposed method in a real-world application with missing data.

Methods

Optimal clustering

Given a point set $S \subset \mathbb{R}^d$, where d is the dimension of the space, denote the number of points in S by $\eta(S)$. A *random labeled point process* (RLPP) is a pair (Ξ, Λ), where Ξ is a point process generating S and Λ generates random labels on point set S. Ξ maps from a probability space to $[N; \mathcal{N}]$, where N is the family of finite sequences in $\mathbb{R}^d$ and $\mathcal{N}$ is the smallest σ-algebra on N such that for any Borel set B in $\mathbb{R}^d$, the mapping $S \to \eta(S \cap B)$ is measurable. A random labeling is a family, $\Lambda = \{\Phi_S : S \in N\}$, where Φ_S is a random label function on the point set S in N. Denoting the set of labels by $L = \{1, 2, ..., l\}$, Φ_S has a probability mass function on L^S defined by $P_S(\phi_S) = P(\Phi_S = \phi_S | \Xi = S)$, where $\phi_S : S \to L$ is a deterministic function assigning a label to each point in S.

A label operator λ maps point sets to label functions, $\lambda(S) = \phi_{S,\lambda} \in L^S$. For any set S, label function ϕ_S and label operator λ, the *label mismatch error* is defined as

$$\epsilon_\lambda(S, \phi_S) = \frac{1}{\eta(S)} \sum_{x \in S} I_{\phi_S(x) \neq \phi_{S,\lambda}(x)}, \tag{1}$$

where I_A is an indicator function equal to 1 if A is true and 0 otherwise. The *error of label function* $\lambda(S)$ is computed as $\epsilon_\lambda(S) = \mathbb{E}_{\Phi_S}[\epsilon_\lambda(S, \phi_S)|S]$, and the *error of label operator* λ for the corresponding RLPP is then defined by $\epsilon[\lambda] = \mathbb{E}_\Xi \mathbb{E}_{\Phi_\Xi}[\epsilon_\lambda(\Xi, \phi_\Xi)]$.

Clustering involves identifying partitions of a point set rather than the actual labeling, where a partition of S into l clusters has the form $\mathcal{P}_S = \{S_1, S_2, ..., S_l\}$ such that S_i's are disjoint and $S = \bigcup_{i=1}^{l} S_i$. A cluster operator ζ maps point sets to partitions, $\zeta(S) = \mathcal{P}_{S,\zeta}$. Considering the label switching property of clustering operators, let us define F_ζ as the family of label operators that all induce the same partitions as the clustering operator ζ. More precisely, a label function ϕ_S induces partition $\mathcal{P}_S = \{S_1, S_2, ..., S_l\}$, if $S_i = \{x \in S : \phi_S(x) = l_i\}$ for distinct $l_i \in L$. Thereby, $\lambda \in F_\zeta$ if and only if $\phi_{S,\lambda}$ induces the same partition as $\zeta(S)$ for all $S \in N$. For any set S, label function ϕ_S and cluster operator ζ, define the *cluster mismatch error* by

$$\epsilon_\zeta(S, \phi_S) = \min_{\lambda \in F_\zeta} \epsilon_\lambda(S, \phi_S), \tag{2}$$

the *error of partition* $\zeta(S)$ by $\epsilon_\zeta(S) = \mathbb{E}_{\Phi_S}[\epsilon_\zeta(S, \phi_S)|S]$ and the *error of cluster operator* ζ for the RLPP by $\epsilon[\zeta] = \mathbb{E}_\Xi \mathbb{E}_{\Phi_\Xi}[\epsilon_\zeta(\Xi, \phi_\Xi)]$.

As shown in [15], error definitions for partitions can be represented in terms of risk with intuitive cost functions.

Specifically, define $G_{\mathcal{P}_S}$ such that $\phi_S \in G_{\mathcal{P}_S}$ if and only if ϕ_S induces $\mathcal{P}_S$. The error of partition can be expressed as

$$\epsilon_\zeta(S) = \sum_{\mathcal{P}_S \in \mathcal{K}_S} c_S(\zeta(S), \mathcal{P}_S) P_S(\mathcal{P}_S), \tag{3}$$

where $\mathcal{K}_S$ is the set of all possible partitions of S, $P_S(\mathcal{P}_S) = \sum_{\phi_S \in G_{\mathcal{P}_S}} P_S(\phi_S)$ is the probability mass function on partitions $\mathcal{P}_S$ of S, and the *partition cost function* between partitions $\mathcal{P}_S$ and $\mathcal{Q}_S$ of S is defined as

$$c_S(\mathcal{Q}_S, \mathcal{P}_S) = \frac{1}{\eta(S)} \min_{\phi_{S,\mathcal{Q}_S} \in G_{\mathcal{Q}_S}} \sum_{x \in S} I_{\phi_{S,\mathcal{P}_S} \neq \phi_{S,\mathcal{Q}_S}}, \tag{4}$$

with $\phi_{S,\mathcal{P}_S}$ being any member of $G_{\mathcal{P}_S}$. A Bayes cluster operator ζ^* is a clusterer with the minimal error $\epsilon[\zeta^*]$, called the *Bayes error*, obtained by a Bayes partition, $\zeta^*(S)$ for each set $S \in N$:

$$\begin{aligned} \zeta^*(S) &= \arg\min_{\zeta(S) \in \mathcal{K}_S} \epsilon_\zeta(S) \\ &= \arg\min_{\zeta(S) \in \mathcal{K}_S} \sum_{\mathcal{P}_S \in \mathcal{K}_S} c_S(\zeta(S), \mathcal{P}_S) P_S(\mathcal{P}_S). \end{aligned} \tag{5}$$

The Bayes clusterer can be solved for each fixed S individually. More specifically, the search space in the minimization and the set of partitions with known probabilities in the summation can be constrained to subsets of $\mathcal{K}_S$, denoted by $\mathcal{C}_S$ and $\mathcal{R}_S$, respectively. We refer to $\mathcal{C}_S$ and $\mathcal{R}_S$ as the set of candidate partitions and the set of reference partitions, respectively. Following [15], we can search for the optimal clusterer based on both optimal and suboptimal procedures (detailed in "Results and discussion" section) with derived bounds that can be used to optimally reduce the size of $\mathcal{C}_S$ and $\mathcal{R}_S$.

Gaussian model with missing values

We consider an RLPP model that generates the points in the set S according to a Gaussian model, where features of $x \in S$ can be missing completely at random due to a missing data mechanism independent of the RLPP. More precisely, the points $x \in S$ with label $\phi_S(x) = i$ are drawn independently from a Gaussian distribution with parameters $\rho_i = \{\mu_i, \Sigma_i\}$. Assuming n_i sample points with label i, we divide the observations into $G_i \leq n_i$ groups, where all n_{ig} points in group g have the same set, J_{ig}, of observed features with cardinality $|J_{ig}| = d_{ig}$. Denoting by S_{ig} the set of sample points in group g of label i, we represent the pattern of missing data in this group using a $d_{ig} \times d$ matrix M_{ig}, where each row is a d-dimensional vector with a single non-zero element with value 1 corresponding to the observed feature's index. Thus, the non-missing portion of sample point $x \in S_{ig}$, i.e. $M_{ig}x$, has Gaussian distribution $\mathrm{N}(M_{ig}\mu_i, M_{ig}\Sigma_i M_{ig}^T)$.

Given $\rho = \{\rho_1, \rho_2, ..., \rho_l\}$ of independent parameters, to evaluate the posterior probability of random labeling function $\phi_S \in L^S$, we have

$$\begin{aligned} &P_S(\phi_S) \propto P(\phi_S) f(S|\phi_S) = \\ &P(\phi_S) \int f(S|\phi_S, \rho) f(\rho) d\rho = \\ &P(\phi_S) \prod_{\substack{i=1 \\ n_i \geq 1}}^{l} \int \left(\prod_{x \in S_i} f_i(x|\rho_i) \right) f(\rho_i) d\rho_i = \\ &P(\phi_S) \prod_{\substack{i=1 \\ n_i \geq 1}}^{l} \int \Big(\prod_{g=1}^{G_i} \prod_{x \in S_{ig}} \\ &\mathrm{N}\left(M_{ig}x; M_{ig}\mu_i, M_{ig}\Sigma_i M_{ig}^T\right) \Big) f(\mu_i, \Sigma_i) d\mu_i d\Sigma_i, \end{aligned} \tag{6}$$

where $P(\phi_S)$ is the prior probability on label functions, which we assume does not depend on the specific points in S.

Known means and covariances

When mean and covariance parameters of label-conditional distributions are known, the prior probability $f(\mu_i, \Sigma_i)$ in (6) is a point mass at $\rho_i = \{\mu_i, \Sigma_i\}$. Thus,

$$\begin{aligned} &P_S(\phi_S) \propto P(\phi_S) \times \\ &\prod_{\substack{i=1 \\ n_i \geq 1}}^{l} \prod_{g=1}^{G_i} \prod_{x \in S_{ig}} \Big[(2\pi)^{-d_{ig}/2} \left| M_{ig}\Sigma_i M_{ig}^T \right|^{-1/2} \times \\ &\exp\left\{ -\tfrac{1}{2}(x-\mu_i)^T M_{ig}^T \left(M_{ig}\Sigma_i M_{ig}^T \right)^{-1} M_{ig}(x-\mu_i) \right\} \Big]. \end{aligned} \tag{7}$$

We define the group-g statistics of label i as

$$\begin{aligned} m_{ig} &:= \frac{1}{n_{ig}} \sum_{x \in S_{ig}} M_{ig}x, \\ \Psi_{ig} &:= \sum_{x \in S_{ig}} (M_{ig}x - m_{ig})(M_{ig}x - m_{ig})^T, \end{aligned} \tag{8}$$

where m_{ig} and Ψ_{ig} are the sample mean and scatter matrix, employing only the observed n_{ig} data points in group g of label i. The posterior probability of labeling function (7) then can be expressed as

$$\begin{aligned} &P_S(\phi_S) \propto P(\phi_S) \prod_{\substack{i=1 \\ n_i \geq 1}}^{l} \prod_{g=1}^{G_i} \\ &\Big[\left| 2\pi\,\Sigma_{ig} \right|^{-n_{ig}/2} \exp\left\{ -\tfrac{1}{2} \mathrm{tr}\left(\Psi_{ig} (\Sigma_{ig})^{-1} \right) \right\} \times \\ &\exp\left\{ -\tfrac{1}{2} n_{ig} (m_{ig} - M_{ig}\mu_i)^T (\Sigma_{ig})^{-1} (m_{ig} - M_{ig}\mu_i) \right\} \Big], \end{aligned} \tag{9}$$

where $\Sigma_{ig} = M_{ig}\Sigma_i M_{ig}^T$ is the covariance matrix corresponding to group g of label i.

Gaussian means and known covariances

Under this model, data points are generated according to Gaussians whose mean parameters are random and their covariance matrices are fixed. Specifically, for label i we have $\mu_i \sim \mathrm{N}(m_i, \frac{1}{\nu_i}\Sigma_i)$, where $\nu_i > 0$ and m_i is a length d real vector. Thus the posterior of label function given the point set S can be derived as

$$\begin{aligned} P_S(\phi_S) \propto P(\phi_S) \prod_{\substack{i=1 \\ n_i \geq 1}}^{l} \Big[\prod_{g=1}^{G_i} \big[|2\pi\,\Sigma_{ig}|^{-n_{ig}/2} \times \\ \exp\left\{-\tfrac{1}{2}\mathrm{tr}\left(\Psi_{ig}(\Sigma_{ig})^{-1}\right)\right\} \big] \times (\nu_i)^{d/2} |2\pi\,\Sigma_i|^{-1/2} \\ \int \exp\Big\{ -\tfrac{1}{2} \textstyle\sum_{g=1}^{G_i} n_{ig}(m_{ig} - M_{ig}\mu_i)^T (\Sigma_{ig})^{-1} \\ (m_{ig} - M_{ig}\mu_i) - \tfrac{\nu_i}{2}(\mu_i - m_i)^T \Sigma_i^{-1}(\mu_i - m_i) \Big\} d\mu_i \Big]. \end{aligned} \tag{10}$$

By completing the square and using the normalization constant of multivariate Gaussian distribution, the integral in this equation can be expressed as

$$\begin{aligned} \int \exp\Big\{ -\tfrac{1}{2} \Big[(\mu_i - A_i^{-1}b_i)^T A_i (\mu_i - A_i^{-1}b_i) + \\ \textstyle\sum_{g=1}^{G_i} n_{ig} m_{ig}^T \Sigma_{ig}^{-1} m_{ig} + \nu_i m_i^T \Sigma_i^{-1} m_i - b_i^T A_i^{-1} b_i \Big] \Big\} \\ = |A_i/(2\pi)|^{-1/2} \exp\Big\{ -\tfrac{1}{2} \Big[\textstyle\sum_{g=1}^{G_i} n_{ig} m_{ig}^T \Sigma_{ig}^{-1} m_{ig} + \\ \nu_i m_i^T \Sigma_i^{-1} m_i - b_i^T A_i^{-1} b_i \Big] \Big\}, \end{aligned} \tag{11}$$

where

$$A_i = \sum_{g=1}^{G_i} n_{ig} M_{ig}^T \Sigma_{ig}^{-1} M_{ig} + \nu_i \Sigma_i^{-1}, \tag{12}$$

$$b_i = \sum_{g=1}^{G_i} n_{ig} M_{ig}^T \Sigma_{ig}^{-1} m_{ig} + \nu_i \Sigma_i^{-1} m_i. \tag{13}$$

Gaussian-Inverse-Wishart Means and Covariances

Under this model, data points are generated from Gaussian distributions with random mean and covariance parameters. More precisely, the parameters associated with label i are distributed as $\mu_i|\Sigma_i \sim \mathrm{N}\left(m_i, \frac{1}{\nu_i}\Sigma_i\right)$ and $\Sigma_i \sim \mathrm{IW}(\kappa_i, \Psi_i)$, where the covariance has inverse-Wishart distribution

$$f(\Sigma_i) = \frac{|\Psi_i|^{\kappa_i/2}}{2^{\kappa_i d/2}\Gamma_d(\kappa_i/2)} |\Sigma_i|^{\frac{\kappa_i + d + 1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\Psi_i \Sigma_i^{-1}\right)\right). \tag{14}$$

To compute the posterior probability of labeling function (6), we first marginalize out the mean parameters μ_i in a similar fashion to (10), obtaining

$$\begin{aligned} P_S(\phi_S) \propto P(\phi_S) \prod_{\substack{i=1 \\ n_i \geq 1}}^{l} \int \Big[\textstyle\prod_{g=1}^{G_i} |2\pi\,\Sigma_{ig}|^{-n_{ig}/2} \times \\ \exp\left\{ -\frac{1}{2}\mathrm{tr}\left(\Psi_{ig}(\Sigma_{ig})^{-1}\right)\right\} \times \\ (\nu_i)^{d/2} |\Sigma_i|^{-1/2} |A_i/(2\pi)|^{-1/2} \times \\ \exp\Big\{ -\tfrac{1}{2} \Big[\textstyle\sum_{g=1}^{G_i} n_{ig} m_{ig}^T \Sigma_{ig}^{-1} m_{ig} \\ + \quad \nu_i m_i^T \Sigma_i^{-1} m_i - b_i^T A_i^{-1} b_i \Big] \Big\} \Big] f(\Sigma_i) d\Sigma_i. \end{aligned} \tag{15}$$

The integration in the above equation has no closed form solution, thus we resort to Monte Carlo integration for approximating it. Specifically, denoting the term in the brackets in Eq. (15) as $g(\Sigma_i)$, we draw J samples $\Sigma_i^{(j)} \sim \mathrm{IW}(\kappa_i, \Psi_i)$, $j = 1, 2, ..., J$, and then compute the integral as $\frac{1}{J}\sum_{j=1}^{J} g(\Sigma_i^{(j)})$.

Results and discussion

The performance of the proposed method for optimal clustering with missing values at random is compared with some suboptimal versions, two other methods for clustering data with missing values, and also classical clustering algorithms with imputed missing values. The performance comparison is carried out on synthetic data generated from different Gaussian RLPP models with different missing probability setups, and also on a publicly available dataset of breast cancer generated by TCGA Research Network (https://cancergenome.nih.gov/). In our experiments, the results of the exact optimal solution for the RLPP with missing at random (Optimal) is provided for smaller point sets, i.e. wherever computationally feasible. We have also tested two suboptimal solutions, similar to the ideas in [15], for an RLPP with missing at random. In the first one (Subopt. Pmax), the set of reference partitions ($\mathcal{R}_S$) is restricted to a closed ball of a specified radius centered on the partition with the highest probability, where the distance of two partitions is defined as the minimum Hamming distance between labels inducing the partitions. In both Optimal and Pmax, the reference set is further constrained to the partitions that assign the correct number of points to each cluster, but the set of candidate partitions ($\mathcal{C}_S$) includes all the possible partitions of n points, i.e. 2^{n-1}. In the second suboptimal solution (Subopt. Pseed), a local search within Hamming distance at 1 is performed starting from five random initial partitions to approximately find the partition with (possibly local)

maximum probability. Then the sets of reference and candidate partitions are constrained to the partitions with correct cluster sizes with a specified Hamming distance from the found (local) maximum probability partition. The bounds derived in [15] for reducing the set of candidate and reference partitions are used, where applicable, in Optimal, Pseed, and Pmax.

In all scenarios, *k*-POD and fuzzy *c*-means with optimal completion strategy (FCM-OCS) are directly applied to the data with missing values. In the simulations in [24], where FCM-OCS is introduced, to initialize cluster centers, the authors apply ordinary fuzzy *c*-means to the complete data, i.e. using knowledge of the missing values. To have a fair comparison with other methods, we calculate the initial cluster centers for FCM-OCS by applying fuzzy *c*-means to the subset of points with no missing features for lower missing rates. For higher missing rates we impute the missing values by the mean of the corresponding feature values across all points, and then apply fuzzy *c*-means to all the points to initialize the cluster centers. In order to apply the classical algorithms, the missing values are imputed according to [28], by employing a multivariate Gibbs sampler that iteratively generates samples for missing values and parameters based on the observed data. The classical algorithms included in our experiments include *k*-means (KM), fuzzy *c*-means (FCM), hierarchical clustering with single linkage (Hier. (Si)), and hierarchical clustering with complete linkage (Hier. (Co)). Moreover, completely random clusterer (Random) results are also included for performance comparisons.

Simulated data

In the simulation analysis, the number of clusters is fixed at 2, and the dimensionality of the RLPPs (number of features) is set to 5. Additional results for 20 features are provided in Additional file 1. Point generation is done based on a Gaussian mixture model (GMM). Three different scenarios for the parameters of the GMM are considered: *i*) Fixed known means and covariances *ii*) Known covariances and unknown means with Gaussian distributions. *iii*) Unknown means and covariances with Gaussian inverse-Wishart distributions. We select the values of the parameters of the point generation process to have an approximate Bayes error of 0.15. The selected values are shown in Table 1 . For the point set generation, the number of points from each cluster is fixed *a priori*. The distributions are first drawn from the assumed model, and then the points are generated based on the drawn distributions. A subset of the points' features is randomly selected to be hidden based on missing at random with different missing probabilities. Four different setups for the number of points are considered in our simulation analysis: 10 points from each cluster ($n_1 = n_2 = 10$), 12 points from one cluster and 8 points from the other cluster ($n_1 = 12, n_2 = 8$), 35 points from each cluster ($n_1 = n_2 = 35$), and 42 points from one cluster and 28 points from the other cluster ($n_1 = 42, n_2 = 28$). When having unequal sized clusters, in half of the repetitions n_1 points are generated from the first distribution and n_2 points from the second distribution, and vice-versa in the other half. In each simulation repetition, all clustering methods are applied to the points to generate a vector of labels that induces a two-cluster partition. The predicted label vector by each method is compared with the true label vector of each point in the point set to calculate the error of that method on that point set. In other words, for each method the number of points assigned to a cluster different from their true one are counted (after accounting for the label switching issue) and divided by the total number of points ($n = n_1 + n_2$) to calculate the clustering error of that method on the point set. These errors are averaged

Table 1 Parameters for the point generation under three models

Model	Mean vectors	Covariance matrices	Distributions' hyperparameters
Fixed means and covariances	$\mu_1 = 0 \cdot \mathbf{1}_d, \mu_2 = 0.445 \cdot \mathbf{1}_d$	$\Sigma_1 = \Sigma_2 = 0.23 \cdot I_d$	—
Gaussian means and fixed covariances	$\mu_1 \sim \mathrm{N}\left(m_1, \frac{1}{\nu_1}\Sigma_1\right), \mu_2 \sim \mathrm{N}\left(m_2, \frac{1}{\nu_2}\Sigma_2\right)$	$\Sigma_1 = \Sigma_2 = 0.28 \cdot I_d$	$m_1 = 0 \cdot \mathbf{1}_d, m_2 = 0.45 \cdot \mathbf{1}_d,$ $\nu_1 = 30, \nu_2 = 5$
Gaussian means and inverse-Wishart covariances	$\mu_1 \sim \mathrm{N}\left(m_1, \frac{1}{\nu_1}\Sigma_1\right), \mu_2 \sim \mathrm{N}\left(m_2, \frac{1}{\nu_2}\Sigma_2\right)$	$\Sigma_1 \sim \mathrm{IW}(\kappa_1, \Psi_1), \Sigma_2 \sim \mathrm{IW}(\kappa_2, \Psi_2)$	$m_1 = 0 \cdot \mathbf{1}_d, m_2 = 0.45 \cdot \mathbf{1}_d,$ $\nu_1 = 30, \nu_2 = 5,$ $\Psi_1 = \Psi_2 = 20.7 \cdot I_d,$ $\kappa_1 = \kappa_2 = 75$

N, IW, $\mathbf{1}_d$, and I_d denote Gaussian, inverse-Wishart, column vector of all ones with length d, and $d \times d$ idendity matrix, respectively

across all point sets in different repetitions to empirically estimate the clustering error of each method under a model and fixed missing-value probability. In cases with $n = 70$, since applying Optimal and Pmax is computationally prohibitive, we only provide the results for Pseed.

In Additional file 1, the average clustering errors are shown as a function of the Hamming distance threshold used to define the set of reference partitions in Pmax and Pseed, for different simulation scenarios. From the Figures in Additional file 1, we see that in all cases, the performances of Pmax and Pseed are quite insensitive to the set threshold of the Hamming distance for reference partitions. Note that in these types of figures all the other methods' performances other than Pmax and Pseed are constant in each plot.

The average results for the fixed mean vectors and covariance matrices across 100 repetitions are shown in Fig. 1. Here, the Hamming distance threshold for reference partitions in Pmax and Pseed is fixed at 1. It can be seen that Optimal, Pmax, and Pseed outperform all the other methods in all the smaller sample size settings, and Pmax and Pseed have virtually the same performance as Optimal. For the larger sample size settings where only Pseed is applied, its superior performance compared with other methods is clear from the figure.

Figure 2 depicts the comparison results under the unknown mean vectors with Gaussian distributions and fixed covariance matrices averaged over 80 repetitions. The Hamming distance threshold in Pmax and Pseed is set to 2. For smaller sample sizes, Optimal, Pmax and Pseed

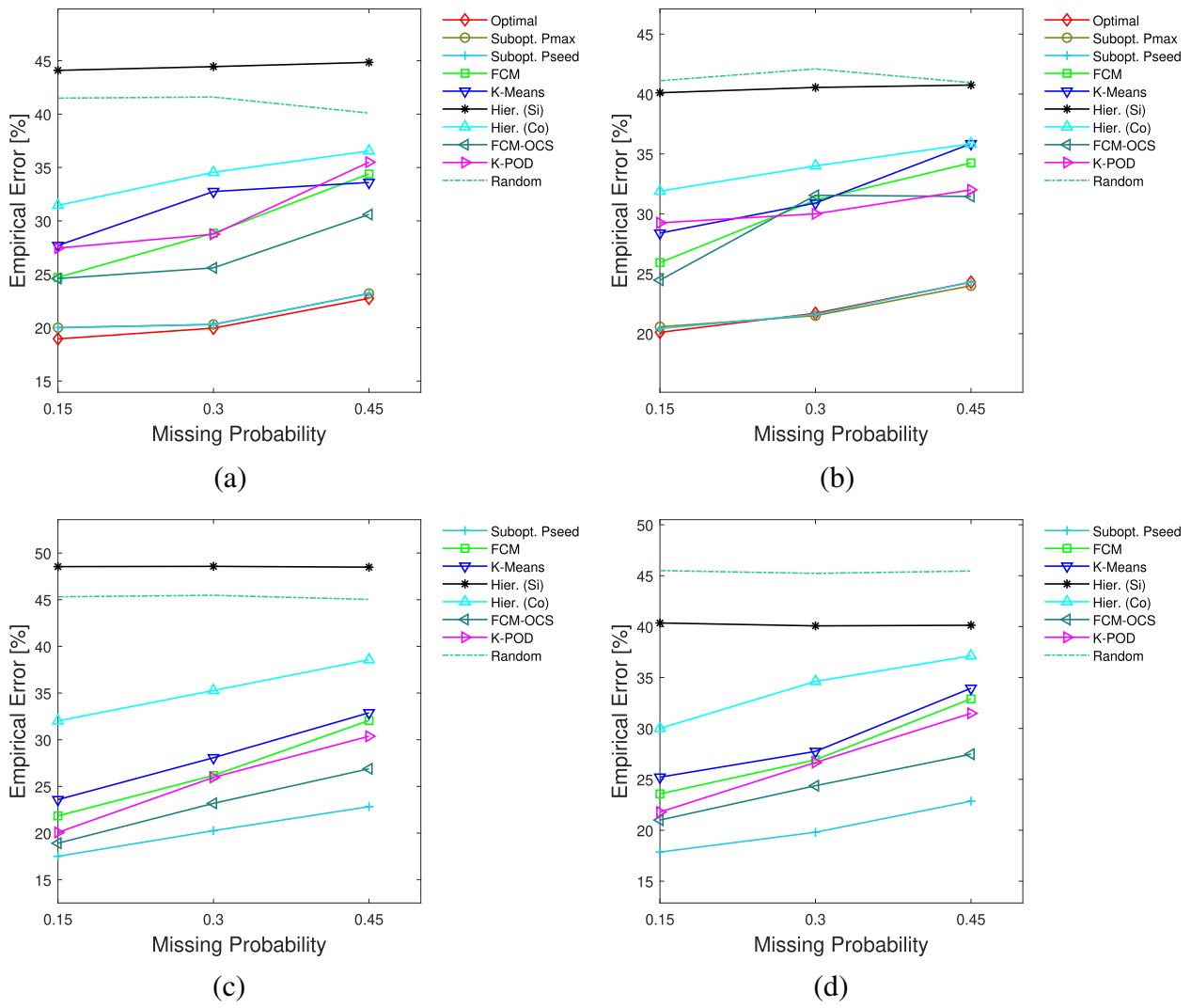

Fig. 1 Average clustering errors vs. missing probability for fixed means and covariances model. The first and second rows correspond to $n = 20$ and $n = 70$, respectively. **a** $n1 = 10, n2 = 10$, **b** $n1 = 12, n2 = 8$, **c** $n1 = 35, n2 = 35$, **d** $n1 = 42, n2 = 28$

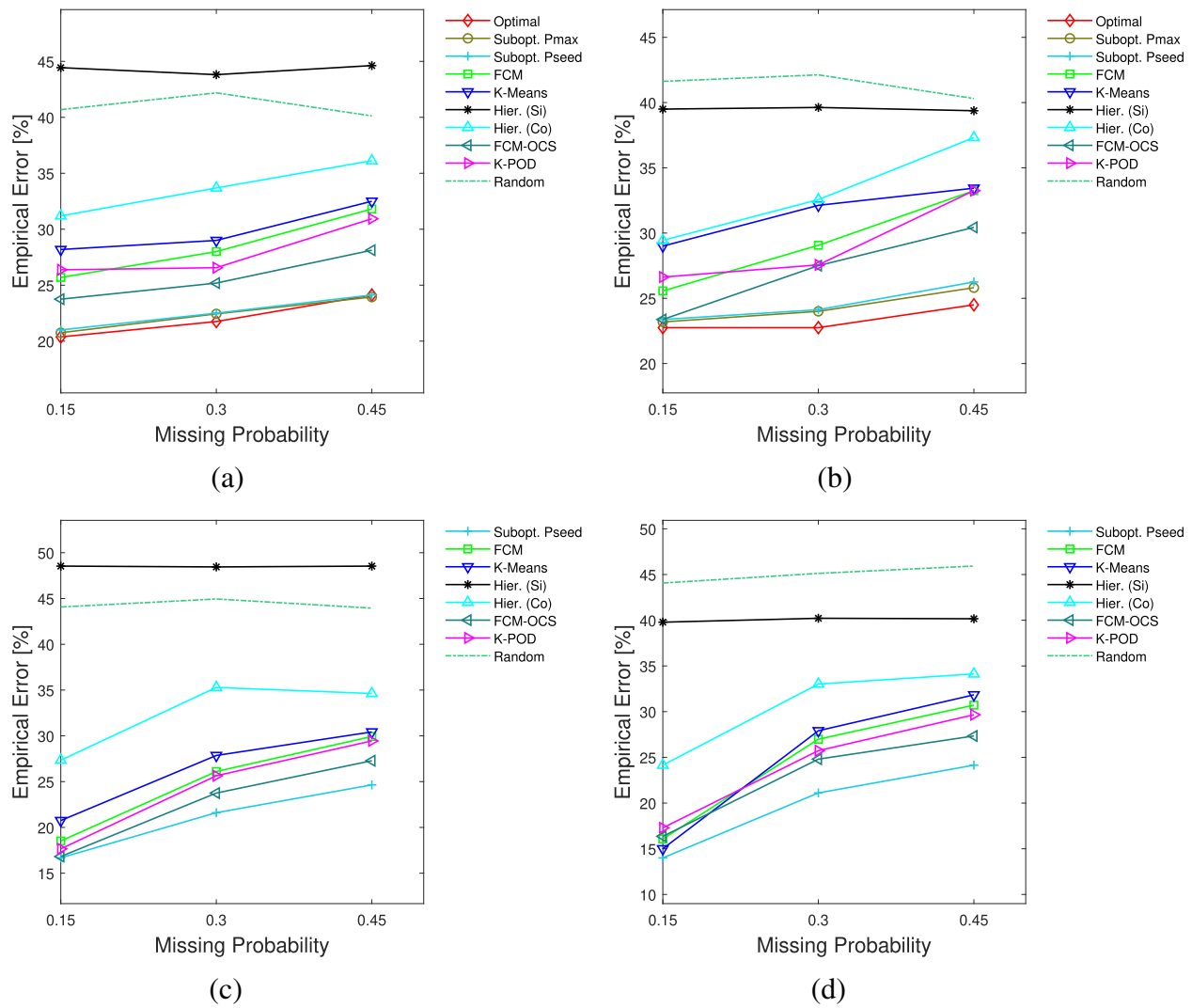

Fig. 2 Average clustering errors vs. missing probability for Gaussian means and fixed covariances model. The first and second rows correspond to $n = 20$ and $n = 70$, respectively. **a** $n1 = 10, n2 = 10$, **b** $n1 = 12, n2 = 8$, **c** $n1 = 35, n2 = 35$, **d** $n1 = 42, n2 = 28$

have lower average errors than all the other methods. We can see that for balanced clusters the suboptimal and optimal solutions have very close performances, but for the unbalanced clusters case with higher missing probabilities the difference between Optimal and Pmax and Pseed gets noticeable. For larger sample sizes Pseed consistently outperforms the other methods, although for lower missing probabilities it has closer competitors. In all cases, as the missing probability increases, the superior performance of the proposed methods becomes more significant.

The average results under the unknown mean vectors and coavriance matrices with Gaussian-inverse-Wishart distribution over 40 repetitions are provided in Fig. 3. In the smaller sample size cases, the Hamming distance threshold in Pmax and Pseed is fixed at 8, and we can see that the proposed suboptimal (Pmax and Pseed) and optimal clustering with missing values have very close average errors, and all are much lower than the other methods' errors. For larger sample sizes, only the results for missing probability equal to 0.15 are shown vs. the Hamming distance threshold used to define the reference partitions in Pseed. Again, Pseed performs better than the other methods.

RNA-seq data

In this section the performance of the clustering methods are examined on a publicly available RNA-seq dataset of breast cancer. The data is available on The Cancer Genome Atlas (TCGA) [29], and is procured by the R package TCGS2STAT [30]. It consists of matched tumor and normal samples, and includes 97 points from each. The original data are in terms of the number of sequence reads mapped to each gene. RNA-seq data are integers, highly skewed and over-dispersed [31–35]. Thus, we apply

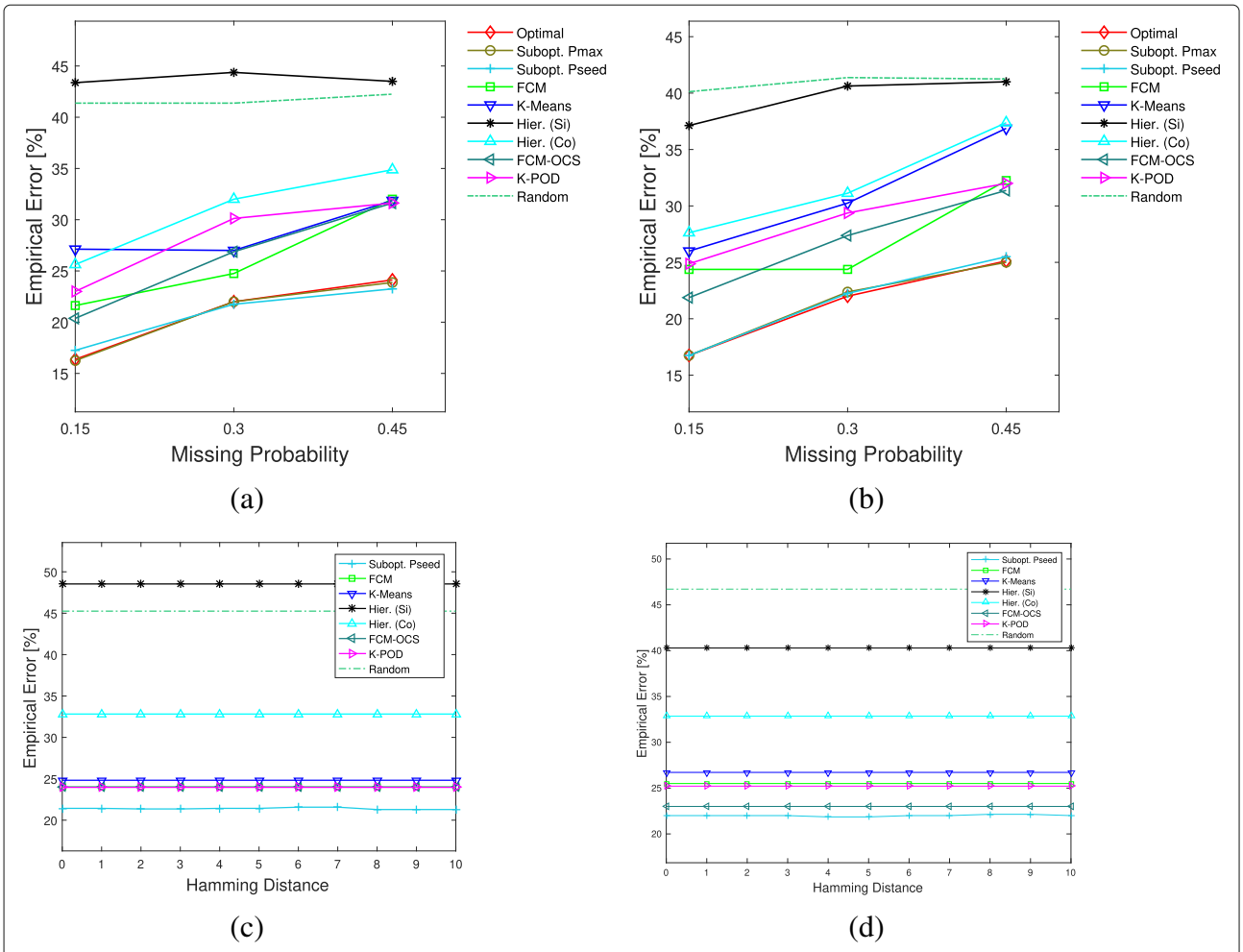

Fig. 3 Average clustering errors for Gaussian means and inverse-Wishart covariances model. The first row corresponds to $n = 20$, and the errors are shown for different missing probabilities. The second row corresponds to $n = 70$ and missing probability of 0.15, where the errors are plotted vs. the Hamming distance threshold used to define the reference partitions in Pseed. **a** $n1 = 10, n2 = 10$, **b** $n1 = 12, n2 = 8$, **c** $n1 = 35, n2 = 35$, miss. prob. $= 0.15$, **d** $n1 = 42, n2 = 28$, miss. prob. $= 0.15$

a variance stabilizing transformation (VST) [36] implemented in DESeq2 package [37], and transform data to a log2 scale that have been normalized with respect to library size. For all subsequent analysis, other than for calculating clustering errors, we assume that the labels of data are unknown. Feature selection is performed in a completely unsupervised manner, since in clustering no labels are known in practice. The top 10 genes in terms of variance to mean ratio of expression are picked as features to be used in clustering algorithms. In general, for setting prior hyperparameters, external sources of information like signaling pathways, where available, can be leveraged [38, 39]. Here, we only use a subset of the discarded gene expressions, i.e. the next 90 top genes (in terms of variance to mean ratio of expression), for prior hyperparameters calibration for the optimal/suboptimal approaches. We follow the approach in [40] and employ the method of moments for prior calibration, but unlike [40], a single set of hyperparameters is estimated and used for both clusters, since the labels of data are not available. It is well known that in small sample size settings, estimation of covariance matrices, scatter matrices and even mean vectors may be problematic. Therefore, similar to [40], we assume the following structure

$$\Psi_0 = \Psi_1 = \begin{bmatrix} \sigma^2 & \rho\sigma^2 & \dots & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \dots & \rho\sigma^2 \\ \vdots & \vdots & \ddots & \vdots \\ \rho\sigma^2 & \dots & \dots & \sigma^2 \end{bmatrix}_{d\times d},$$

$$m_0 = m_1 = m[1, \cdots, 1]_d^T,$$

$$\nu_0 = \nu_1 = \nu, \kappa_0 = \kappa_1 = \kappa,$$

and estimate five scalars $(m, \sigma^2, \rho, \kappa, \nu)$ from the data.

In each repetition, stratified sampling is done, i.e. n_1 and n_2 points are sampled randomly from each group (normal and tumor). When $n_1 \neq n_2$, in half of the repetitions n_1 and n_2 points are randomly selected from the normal and tumor samples, respectively, and vice-versa in the other half. Prior calibration is performed in each repetition, and 15% of the selected features are considered as missing values. Similar to the experiments on the simulated data, the clustering error of each method in each iteration is calculated by comparing the predicted labels and true labels of the sampled points (accounting for label switching issue), and the average results over 40 repetitions are provided in Fig. 4. It can be seen that the proposed optimal clustering with missing values and its suboptimal versions outperform the other algorithms. It is worth noting that the performance of Pseed is more sensitive to the selected Hamming distance threshold for reference partitions compared with the results on simulated data.

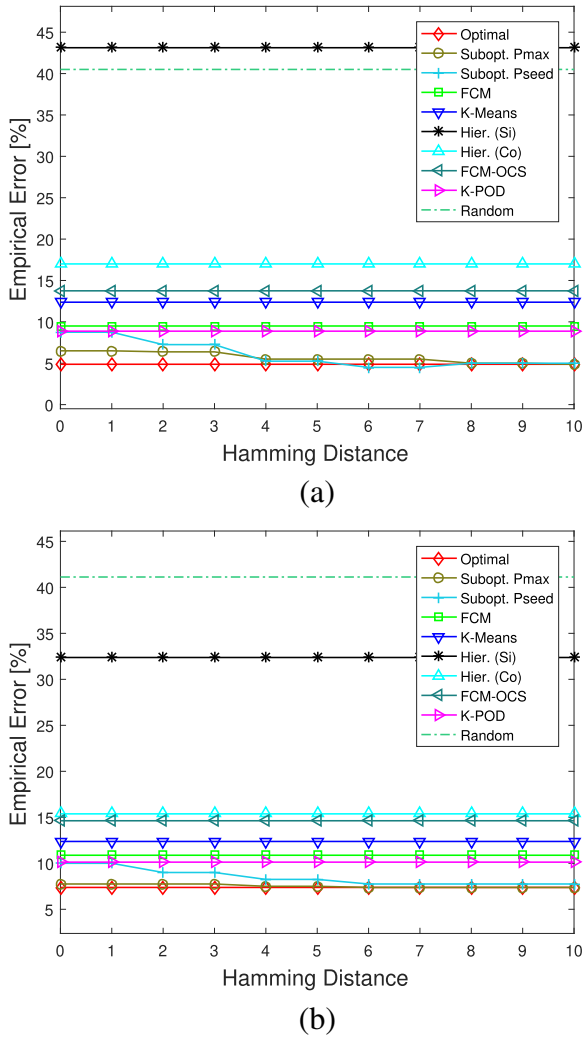

Fig. 4 Empirical clustering errors on breast cancer RNA-seq data. **a** $n1 = 10, n2 = 10$, miss. prob. $= 0.15$, **b** $n1 = 12, n2 = 8$, miss. prob. $= 0.15$

Conclusion

The methodology employed in this paper is very natural. As with any signal processing problem, the basic problem is to find an optimal operator from a class of operators given the underlying random process and a cost function, which is often an error associated with operator performance. While it may not be possible to compute the optimal operator, one can at least employ suboptimal approximations to it while knowing the penalties associated with the approximations.

In this paper, we have, in effect, confronted an old problem in signal processing: If we wish to make a decision based on a noisy observed signal, is it better to filter the observed signal and then determine the optimal decision on the filtered signal, or to find the optimal decision based directly on the observed signal? The answer is the latter. The reason is that the latter approach is fully optimal relative to the actual observation process, whereas, even if in the first approach the filtering is optimal relative to the noise process, the first approach produces a composite of two actions, filter and decision, each of which is only optimal relative to a portion of the actual observation process. In the present situation involving clustering, in the standard imputation-followed-by-clustering approach, it is typically the case that neither the filter (imputation) nor the decision (clustering) is optimal, so that even more advantage is obtained by optimal clustering over the missing-value-adjusted RLPP.

Additional file

Additional file 1: Supplementary materials. Additional figures are given in a single multi-page PDF. (PDF 516 KB)

Abbreviations

FCM-OCS: Fuzzy *c*-means with optimal completion strategy; GMM: Gaussian mixture model; Hier. (Co): Hierarchical clustering with complete linkage; Hier. (Si): Hierarchical clustering with single linkage; MCAR: Missing completely at random; RLPP: Random labeled point process; RNA-seq: RNA sequencing; TCGA: The Cancer Genome Atlas; VST: Variance stabilizing transformation

Acknowledgment

We thank Texas A&M High Performance Research Computing for providing computational resources to perform experiments in this paper.

Funding

This work was funded in part by Awards CCF-1553281 and IIS-1812641 from the National Science Foundation, and a DMREF grant from the National Science Foundation, award number 1534534. The publication cost of this article was funded by Award IIS-1812641 from the National Science Foundation.

Availability of data and materials

The publicly available real datasets analyzed during the current study have been generated by the TCGA Research Network https://cancergenome.nih.gov/.

About this supplement

This article has been published as part of *BMC Bioinformatics Volume 20 Supplement 12, 2019:* Selected original research articles from the Fifth

International Workshop on Computational Network Biology: Modeling, Analysis and Control (CNB-MAC 2018): Bioinformatics. The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-20-supplement-12.

Authors' contributions
SB and SZD developed the method, performed the experiments, and wrote the first draft. XQ and ERD proofread and edited the manuscript, and oversaw the project. All authors have read and approved final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Published: 20 June 2019

References

1. Ben-Dor A, Shamir R, Yakhini Z. Clustering gene expression patterns. J Comput Biol. 1999;6:281–97. https://doi.org/10.1089/106652799318274. PMID: 10582567.
2. Bittner M, Meitzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A, Sampas N, Dougherty E, Wang E, Marincola F, Gooden C, Lueders J, Glatfelter A, Pollock P, Carpten J, Gillanders E, Leja D, Dietrich K, Beaudry C, Berens M, Alberts D, Sondak V, Hayward N, Trent J. Molecular classification of cutaneous malignant melanoma by gene expression profiling. J Comput Biol. 2000;406(3):536–40.
3. Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL. Model-based clustering and data transformations for gene expression data. Bioinformatics. 2001;17(10):977–87.
4. Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. J Am Stat Assoc. 2002;97(458):611–31.
5. MacEachern SN, Müller P. Estimating mixture of Dirichlet process models. J Comput Graph Stat. 1998;7(2):223–38.
6. Dalton LA, Dougherty ER. Optimal classifiers with minimum expected error within a Bayesian framework-part i: Discrete and Gaussian models. Pattern Recogn. 2013;46(5):1301–14. https://doi.org/10.1016/j.patcog.2012.10.018.
7. Imani M, Braga-Neto UM. Control of gene regulatory networks with noisy measurements and uncertain inputs. IEEE Trans Control Netw Syst. 2018;5(2):760–9. https://doi.org/10.1109/TCNS.2017.2746341.
8. Karbalayghareh A, Braga-Neto U, Dougherty ER. Intrinsically Bayesian robust classifier for single-cell gene expression trajectories in gene regulatory networks. BMC Syst Biol. 2018;12(3):23.
9. Imani M, Braga-Neto U. Control of gene regulatory networks using Bayesian inverse reinforcement learning. IEEE/ACM Trans Comput Biol Bioinforma. 20181. https://doi.org/10.1109/TCBB.2018.2830357.
10. Boluki S, Qian X, Dougherty ER. Experimental design via generalized mean objective cost of uncertainty. IEEE Access. 2019;7:2223–30. https://doi.org/10.1109/ACCESS.2018.2886576.
11. Broumand A, Esfahani MS, Yoon B.-J., Dougherty ER. Discrete optimal Bayesian classification with error-conditioned sequential sampling. Pattern Recognit. 2015;48(11):3766–82. https://doi.org/10.1016/j.patcog.2015.03.023.
12. Talapatra A, Boluki S, Duong T, Qian X, Dougherty E, Arróyave R. Autonomous efficient experiment design for materials discovery with Bayesian model averaging. Phys Rev Mater. 2018;2:113803. https://doi.org/10.1103/PhysRevMaterials.2.113803.
13. Karbalayghareh A, Qian X, Dougherty ER. Optimal Bayesian transfer learning. IEEE Trans Signal Process. 2018;66(14):3724–39. https://doi.org/10.1109/TSP.2018.2839583.
14. Dougherty ER, Brun M. A probabilistic theory of clustering. Pattern Recogn. 2004;37(5):917–25. https://doi.org/10.1016/j.patcog.2003.10.003.
15. Dalton LA, Benalcázar ME, Brun M, Dougherty ER. Analytic representation of Bayes labeling and Bayes clustering operators for random labeled point processes. IEEE Trans Signal Process. 2015;63(6):1605–20. https://doi.org/10.1109/TSP.2015.2399870.
16. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995;270(5235):467–70.
17. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-seq. Nat Methods. 2008;5(7):621.
18. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001;17(6):520–5.
19. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. J Stat Softw Artic. 2011;45(3):1–67. https://doi.org/10.18637/jss.v045.i03.
20. Honaker J, King G, Blackwell M, et al. Amelia ii: A program for missing data. J Stat Softw. 2011;45(7):1–47.
21. Stekhoven DJ, Bühlmann P. Missforest—non-parametric missing value imputation for mixed-type data. Bioinformatics. 2011;28(1):112–8.
22. Little RJ, Rubin DB. Statistical Analysis with Missing Data vol. 333. New Jersey: Wiley; 2014.
23. Chi JT, Chi EC, Baraniuk RG. k-pod: A method for k-means clustering of missing data. Am Stat. 2016;70(1):91–9. https://doi.org/10.1080/00031305.2015.1086685.
24. Hathaway RJ, Bezdek JC. Fuzzy c-means clustering of incomplete data. IEEE Trans Sys Man Cybern B (Cybern). 2001;31(5):735–44. https://doi.org/10.1109/3477.956035.
25. Kanungo T, Mount DM, Netanyahu NS, Piatko CD, Silverman R, Wu AY. An efficient k-means clustering algorithm: Analysis and implementation. IEEE Trans Pattern Anal Mach Intell. 2002;24(7):881–92.
26. Bezdek JC, Ehrlich R, Full W. FCM: The fuzzy c-means clustering algorithm. Comput Geosci. 1984;10(2-3):191–203.
27. Johnson SC. Hierarchical clustering schemes. Psychometrika. 1967;32(3):241–54.
28. Dadaneh SZ, Dougherty ER, Qian X. Optimal Bayesian classification with missing values. IEEE Trans Signal Process. 2018;66(16):4182–92.
29. The Cancer Genome Atlas Research Network (TCGA). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008;455(7216):1061.
30. Wan Y.-W., Allen GI, Liu Z. TCGA2STAT: simple TCGA data access for integrated statistical analysis in R. Bioinformatics. 2015;32(6):952–4.
31. Dadaneh SZ, Zhou M, Qian X. Bayesian negative binomial regression for differential expression with confounding factors. Bioinformatics. 2018;1:. https://doi.org/10.1093/bioinformatics/bty330.
32. Hajiramezanali E, Zamani Dadaneh S, Karbalayghareh A, Zhou M, Qian X. Bayesian multi-domain learning for cancer subtype discovery from next-generation sequencing count data. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, editors. Advances in Neural Information Processing Systems 31. Curran Associates, Inc.; 2018. p. 9115–24.
33. Dadaneh SZ, Qian X, Zhou M. BNP-Seq: Bayesian nonparametric differential expression analysis of sequencing count data. J Am Stat Assoc. 2018;113(521):81–94.
34. Hajiramezanali E, Dadaneh SZ, de Figueiredo P, Sze S.-H., Zhou M, Qian X. Differential expression analysis of dynamical sequencing count data with a Gamma Markov chain. 2018. arXiv preprint arXiv:1803.02527.
35. Broumand A, Hu T. A length bias corrected likelihood ratio test for the detection of differentially expressed pathways in RNA-seq data. In: 2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP); 2015. p. 1145–9. https://doi.org/10.1109/GlobalSIP.2015.7418377.
36. Durbin BP, Hardin JS, Hawkins DM, Rocke DM. A variance-stabilizing transformation for gene-expression microarray data. Bioinformatics. 2002;18(suppl_1):105–10.
37. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
38. Boluki S, Esfahani MS, Qian X, Dougherty ER. Constructing Pathway-Based Priors within a Gaussian Mixture Model for Bayesian Regression and Classification. IEEE/ACM Trans Comput Biol Bioinforma. 2019;16(2):524–37. https://doi.org/10.1109/TCBB.2017.2778715. ISSN: 1545-5963.
39. Boluki S, Esfahani MS, Qian X, Dougherty ER. Incorporating biological prior knowledge for Bayesian learning via maximal knowledge-driven information priors. BMC Bioinformatics. 2017;18(14):552. https://doi.org/10.1186/s12859-017-1893-4.
40. Dalton LA, Dougherty ER. Application of the Bayesian MMSE estimator for classification error to gene expression microarray data. Bioinformatics. 2011;27(13):1822–31.