Journal of Bioinformatics and Computational Biology
Vol. 21, No. 4 (2023) 2350019 (26 pages)
© World Scientific Publishing Europe Ltd.
DOI: 10.1142/S0219720023500191

World Scientific
www.worldscientific.com

A model-based clustering algorithm with covariates adjustment and its application to lung cancer stratification

Carlos E. M. Relvas[*,¶], Asuka Nakata[†,‖], Guoan Chen[‡,**], David G. Beer[§,††],
Noriko Gotoh[†,‡‡] and Andre Fujita[*,§§]

[*] *Institute of Mathematics and Statistics, University of São Paulo*
Rua do Matão 1010 São Paulo, São Paulo 05508-090, Brazil
[†] *Cancer Research Institute, Kanazawa University*
Kanazawa, Ishikawa 920-1164, Japan
[‡] *School of Medicine, Southern University of Science and Technology,*
1088 Xueyuan Blvd. Shenzhen, Guangdong 518055, P .R. China
[§] *Rogel Cancer Center, University of Michigan,*
1500 E Medical Center Dr Ann Arbor, Michigan 48109, USA
[¶] *carlos.edu.relvas@gmail.com*
[‖] *a.nakata.bs@gmail.com*
[**] *cheng@sustech.edu.cn*
[††] *dgbeer@med.umich.edu*
[‡‡] *ngotoh@staff.kanazawa-u.ac.jp*
[§§] *andrefujita@usp.br*

Received 24 March 2022
Revised 10 May 2023
Accepted 30 July 2023
Published 8 September 2023

Usually, the clustering process is the first step in several data analyses. Clustering allows identify patterns we did not note before and helps raise new hypotheses. However, one challenge when analyzing empirical data is the presence of covariates, which may mask the obtained clustering structure. For example, suppose we are interested in clustering a set of individuals into controls and cancer patients. A clustering algorithm could group subjects into young and elderly in this case. It may happen because the age at diagnosis is associated with cancer. Thus, we developed CEM-Co, a model-based clustering algorithm that removes/minimizes undesirable covariates' effects during the clustering process. We applied CEM-Co on a gene expression dataset composed of 129 stage I non-small cell lung cancer patients. As a result, we identified a subgroup with a poorer prognosis, while standard clustering algorithms failed.

Keywords: Mixture Gaussian models; EM algorithm; clustering; lung cancer.

[§§] Corresponding author.

1. Introduction

Clustering is commonly the first step in several data analyses. By clustering similar items into the same group, we can identify outliers and patterns we did not note before and suggest new hypotheses. There are several clustering methods in the literature, for example, k-means,[20] k-medoids,[19] hierarchical clustering,[29] model-based clustering,[23] density-based spatial clustering of applications with noise,[14] mean-shift,[8] c-means,[3] and spectral.[28] One challenge when we analyze empirical data is the presence of covariates. We illustrate this case in Fig. 1. Suppose there are two groups, i.e. controls and patients. Due to strong covariates effects, such as age at diagnosis, some individuals of the control group (red dots) are assigned to the patients' group (black dots) and vice versa (Fig. 1(a)). The clustering algorithm separated the items into young and elderly in this case. However, we wished to obtain one group of controls and another of the patients (Fig. 1(b)).

There are several approaches to tackling the clustering problem by minimizing the covariates' effects. One is the one-step latent class analysis (LCA)-based method. The one-step procedure removes the covariates' effects from each item's probability belonging to each cluster.[2,9] However, suppose the strength of the associations between latent classes and their indicators is weak. In that case, auxiliary variables may be affected by the first-phase LCA model's parameters.[1,18,24,26] An alternative solution for this problem would be the three-step approach. The three-step procedure uses a mixture model to cluster the data and logistic regression analysis to infer the associations between the predicted class membership and covariates.[15] The drawback is that it considers only categorical items. Other approaches basis on latent profile analysis, which examines the items as continuous data. However, they assume local

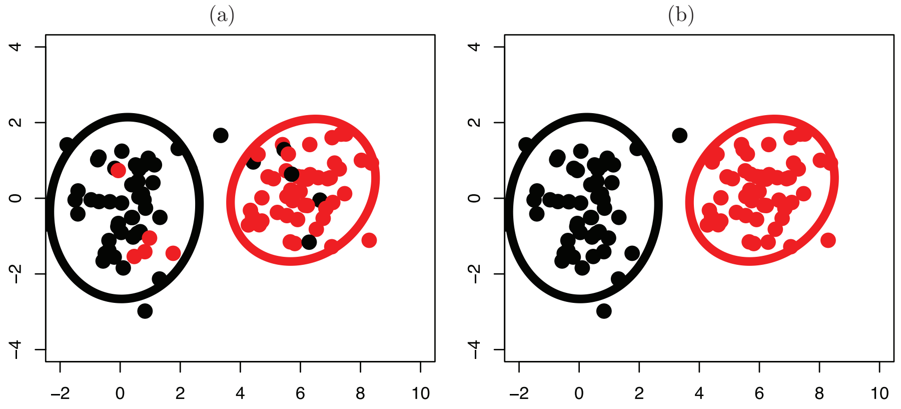

Fig. 1. Schema to illustrate the covariate effect on the clustering structure. Suppose two populations of subjects sampled from normal distributions with means $(-1, -1)$ and $(1, 7)$ and identity covariance matrices. Moreover, consider a covariate effect on the clustering centroids, e.g. age, as follows: $(0.1, 0.1)$ on the first cluster and $(-0.2, -0.05)$ on the second cluster. In this case, whether we do not consider the covariate effect during the clustering process, we cluster some "red" items together with the "black" items and vice versa. We expect to obtain the clustering structure illustrated in panel (b) by considering the covariate effect.

independence and, therefore, a specific covariance matrix structure.[4,27] Moreover, they model the covariate effect on each item's probability belonging to each cluster but not on the cluster's centroid, where we expect the covariate effect. Therefore, we cannot obtain the clustering structure filtered out by the impact of the covariates.

Usually, methods to obtain the clustering structure without covariates' effects consider the covariates as additional dimensions of the items. Alternatively, they remove the covariates' effects before clustering by conducting a linear regression. In both cases, the methods assume that the covariates affect the clusters equally, which may not be valid. For example, age and gender are associated with cancer in different manners.

In this context, we developed a framework, namely CEM-Co, which is composed of the following:

(1) A model-based clustering algorithm that considers the covariates (on the centroids and/or covariances),
(2) A statistical test to identify which covariates are associated with the clustering structure, and
(3) A Bayesian information criterion (BIC) to estimate the number of clusters.

Finally, we applied CEM-Co in non-small cell lung cancer (NSCLC) stage I gene expression data.

2. Methods

Let *N, K*, and *P* be the number of items, clusters, and covariates, respectively. Also, let $\mathbf{x}_i$ be the *i*th *M*-dimensional item ($i = 1,\ldots,N$), and $\mathbf{z}_i = (\mathbf{z}_{1,i}, \mathbf{z}_{2,i},\ldots,\mathbf{z}_{P,i})$ be the covariates associated with $\mathbf{x}_i$. Then, our goal is to cluster the *N* items $(\mathbf{x}_1, \mathbf{x}_2,\ldots,\mathbf{x}_N)$. into *K* clusters by considering the *P* covariates $(\mathbf{z}_1, \mathbf{z}_2,\ldots,\mathbf{z}_p)$.

2.1. *CEM-Co*

Like the classification expectation–maximization (EM) algorithm,[6,7] CEM-Co represents each cluster by a normal distribution. Therefore, a mixture of normal distributions can model the entire dataset.

Let μ_j^* ($j = 1,\ldots,K$) be the clusters' centroids (means of the normal distributions) without the covariates' effects, $\beta_{j,i}$ be the *M*-dimensional coefficient representing the strength of the *l*th covariate effect ($l = 1,\ldots,P$) on the *j*th cluster centroid. Then, we model the "observed" clusters centroids as follows:

$$\mu_{i,j}|\mathbf{z}_{\mathrm{i}} = \mu_j^* + \beta_{j,1}\mathbf{z}_{1,\mathrm{t}} + \beta_{j,2}\mathbf{z}_{2,\mathrm{i}} + \cdots + \beta_{j,P}\mathbf{z}_{P,\mathrm{i}}. \tag{1}$$

Let $\mathbf{L}_{i,j}$ be a diagonal $M \times M$ matrix whose *r*th diagonal ($r = 1,\ldots,M$) is $\sigma_{r,j} + \gamma_{1,r,j}\mathbf{z}_{1,i} + \gamma_{2,r,j}\mathbf{z}_{2,i} +\ldots+ \gamma_{P,r,j}\mathbf{z}_{P,i}$, and $\mathbf{E}_j$ be an $M \times M$ positive-definite matrix. Then, we model the covariance matrix of the multivariate normal distribution (cluster) as follows:

$$\mathbf{\Sigma}_{i,j} = \mathbf{L}_{i,j}\mathbf{E}_{j}\mathbf{L}_{i,j}. \tag{2}$$

The $\mathbf{\Sigma}_{i,j}$ matrix is a Cholesky decomposition[11] to guarantee a positive-definite matrix and accommodate the covariate effect in the covariance structure. Suppose there is no covariate effect on the $\mathbf{\Sigma}_{i,j}$ covariate matrix. Then, $\sigma_{r,j} = 1$ and $\gamma_{l,r,j} = 0$ for all $l = 1,\ldots,P$, $r = 1,\ldots,M$, and $j = 1,\ldots,K$. In other words, $\mathbf{L}_{i,j}$ is an identity matrix. Consequently, the covariance matrix is defined only by $\mathbf{E}_j$.

Let

$$\phi\left(\mathbf{x}_i, \mu_{i,j}, \mathbf{\Sigma}_{i,j}\right) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}_1 - \mu_{i,j})^{\top}\mathbf{\Sigma}_{i,j}^{-1}(\mathbf{x}_i - \mu_{i,j})\right)}{\sqrt{\det(2\pi\mathbf{\Sigma}_{i,j})}}, \tag{3}$$

be the density of a multivariate normal distribution, and α_j be the weight associated with the jth cluster $\left(\sum_{j-1}^{K} \alpha_j = 1\right)$. Then, we can write the likelihood function as follows:

$$L(\theta \mid \mathbf{x}_1, \ldots, \mathbf{x}_N) = \prod_{i=1}^{N}\sum_{j=1}^{K}\alpha_j\phi\left(\mathbf{x}_{\mathrm{i}}, \mu_{i,j}, \mathbf{\Sigma}_{i,j}\right), \tag{4}$$

where $\theta = (\alpha, \mu^{*}, \beta, \mathbf{L}, \mathbf{E})$. Based on Eqs. (1)–(4), we can simultaneously estimate the parameters (μ_j^{*}, $\mathbf{E}_j$, $\sigma_{r,j}$, $\gamma_{l,r,j}$, $\beta_{j,l}$, α_j). Also, we can carry out the clustering by using a variant of the EM algorithm. In other words, we make hard assignments to the latent variables. We describe the CEM-Co in Algorithm 1.

Similar to the classification EM algorithm, CEM-Co depends on the initialization of the parameters in step 2 of Algorithm 1 (α_j, μ_j^{*}, $\beta_{j,l}$, $\mathrm{L}_{i,j}$, E_j, for $i = 1,\ldots,N$, $j = 1,\ldots,K$, and $l = 1,\ldots,P$). Different initializations may lead to different likelihood values (the EM algorithm usually obtains a local optimum).

Thus, we constructed a linear regression model for each feature in the cluster ($\mathbf{x}$). The feature represents the response variable, and the covariates ($\mathbf{z}$) represent the explanatory variables. Then, we run the clustering algorithm using the residuals of all regressions as our new cluster space. We describe the initialization of the parameters in Algorithm 2.

We obtain different initialization values by varying the parameter f of Algorithm 2. If we set $f = 0$, we obtain the standard EM algorithm. In all simulations, we used 20 different initializations by varying f from 0 to 2 ($f = 0$, 0.1, 0.2,...,1.9, 2). Note that, the same covariate can affect the mean and the covariance structure. Also, we assume that covariates always enter all components when they are in the model.

Algorithm 1. Classification EM with covariates effects on both the clusters centroids and covariance matrices.
Input: the items x_i ($i = 1,\ldots,N$), the number of clusters K, and the covariates $\mathbf{z}_{l,i}$ ($l = 1,\ldots,P$).
Output: the K clusters.

1 Let $\mu_{i,j}$ (Eq. (1)), $\Sigma_{i,j}$ (Eq. (2)), α_j, and $\phi(\mathrm{x}_I, \mu_{i,j}, \Sigma_{i,j})$ (Eq. (3)) be the centroid, the covariance matrix, the weight associated with the jth normal distribution ($j = 1,\ldots,K$), and the density of the multivariate normal distribution, respectively.

2 Randomly initialize the parameters of the K normal distributions, i.e. the centroids (μ_j^*), covariances matrices ($\mathbf{E}_j$, $\sigma_{r,j}$, $\gamma_{l,r,j}$), covariates effects ($\beta_{j,l}$), and weights for each normal distribution (α_j) ($r = 1,\ldots,M$).

3 *Expectation step.* Compute the expected probability for the ith item to belong to the jth cluster as

$$P_{i,j} = \frac{\hat{\alpha}_j \phi(\mathbf{x}_i, \hat{\mu}_{i,j}, \hat{\Sigma}_{i,j})}{\sum_{l=1}^{k} \hat{\alpha}_j \phi(\mathbf{x}_i, \hat{\mu}_{i,j}, \hat{\Sigma}_{i,j})}.$$

4 *Maximization step.* Let $\hat{d}_{1,j} = \hat{\beta}_{j,1}\mathbf{z}_{1,1} + \hat{\beta}_{j,2}\mathbf{z}_{2,1} + \cdots + \hat{\beta}_{j,P}\mathbf{z}_{P,1}$ be the sum of all covariates effects. Then, compute the maximum likelihood estimates for

$$\hat{\mu}_j^* = \frac{\sum_{i=1}^{N}\left(\mathbf{x}_i - \hat{d}_{i,j}\right)\hat{P}_{i,1}}{\sum_{i=1}^{N}\hat{P}_{i,j}}, \quad \hat{\mathbf{E}}_j = \frac{\sum_{i=1}^{N}\hat{P}_{i,j}\left(\hat{\mathbf{L}}_{i,j}^{-1}\left(\mathbf{x}_i - \hat{\mu}_j^* - \hat{d}_{i,j}\right)\right)\left(\hat{\mathbf{L}}_{i,1}^{-1}\left(\mathbf{x}_i - \hat{\mu}_j^* - \hat{d}_{i,j}\right)\right)^{\top}}{\sum_{i=1}^{N}\hat{P}_{i,j}},$$

$$\hat{\beta}_{j,l} = \frac{\sum_{i=1}^{N}\mathbf{z}_{l,i}\left(\mathbf{x}_i - \hat{\mu}_j^* - \hat{d}_{i,j} + \hat{\beta}_{j,l}\mathbf{z}_{l,i}\right)\hat{P}_{i,j}}{\sum_{i=1}^{N}\mathbf{z}_{l,i}^2\hat{P}_{i,j}}, \quad \text{and} \quad \hat{\alpha}_j = \frac{\sum_{i=1}^{N}\hat{P}_{i,j}}{N}.$$

To estimate the maximized value of the likelihood function, find the roots (or zeros) of the partial derivative of $\hat{\mathbf{L}}_{i,j}$ by using the Newton–Raphson method.

5 Go to the Expectation step until convergence (difference in log-likelihoods between iterations less than 0.001) of the likelihood function (Eq. (4))

Algorithm 2. Initialization.
Input: the n items, the number of clusters k, the cluster space $\mathbf{X}$, the covariates' space $\mathbf{Z}$, and a factor f.
Output: the initialization of the parameters.

1 Let M be the dimension of X. For $j = 1,\ldots,M$, carry out a linear regression $\mathbf{X}_j = \beta_j\mathbf{Z} + \varepsilon_j$.
2 Multiply each β_j by f.
3 Create a new space using the residuals ($\varepsilon_1,\ldots,\epsilon_m$), using β_j multiplied by f.
4 Run the EM algorithm on this residual space with k clusters.
5 Return β and the EM parameters as our initializing values.

2.2. *Statistical test for the covariate effect*

In some empirical data analyses, one may be interested in testing whether a covariate is statistically associated with the clustering structure. For example, suppose we want to check whether the lth covariate ($\mathbf{z}_l$) does affect the clustering structure. In other words, consider the following statistical test:

H_0: $\beta_{1,l} = \beta_{2,l} = \cdots = \beta_{K,l} = \mathbf{0}$

versus

H_1: At least one $\beta_{j,l} \neq 0$ ($j = 1,\ldots,K$).

To this end, we propose a likelihood ratio test (LRT).[30] Let $L(\alpha_j, \mu_{\mathrm{j}}^{*}, \beta_{j,1}, \mathbf{L}_{i,j}, \mathrm{E}_j|\mathbf{x}_1,\ldots,\mathbf{x}_N, \beta_{1,l} = \cdots = \beta_{K,l} = 0)$ be the likelihood function assuming that the lth covariate has no effect on the jth centroid (the null model), and $L(\alpha_j, \mu_{\mathrm{j}}^{*}, \beta_{j,1}, \mathbf{L}_{i,j}, \mathrm{E}_j|\mathbf{x}_1,\ldots,\mathbf{x}_N)$ be the likelihood function for the alternative model (Eq. (4)). Then, we can define the statistic of the test (D) as follows:

$$D = 2\Big\{\ln\Big(L\Big(\hat{\theta}|\mathbf{x}_1,\ldots,\mathbf{x}_N\Big)\Big) - \ln\Big(L\Big(\hat{\theta}_0|\mathbf{x}_1,\ldots,\mathbf{x}_N\Big)\Big)\Big\}, \tag{5}$$

where $\hat{\theta}_0$ are the coefficients obtained by the maximum likelihood estimator under H_0 ($\beta_{1,l} = \beta_{2,l} = \cdots = \beta_{K,l} = 0$).

To obtain D, carry out the CEM-Co algorithm with (the alternative model) and without (the null model) the lth covariate and compare the two likelihood functions.

The probability distribution of the test statistic D is approximately a chi-squared distribution with degrees of freedom equal to the number of parameters of the alternative model minus the number of parameters of the null model. For example, in our specific test ($H_0 : \beta_{1,l} = \beta_{2,l} = \cdots = \beta_{K,l} = \mathbf{0}$), we have K parameters to be estimated for all α_j, KM for all μ_j^{*}, KMP for all $\beta_{j,l}$, KM^2 for all $\mathbf{E}_j$, and $2KM$ for all $\mathbf{L}_{i,j}$. Thus, the total number of parameters is $K(1 + M\ (3 + P\ ) + M^2)$. To test the effect of a single covariate, we have $K(1 + M\ (2 + P\ ) + M^2)$ parameters under H_0. Therefore, the number of degrees of freedom for this test is KM.

We may use a similar procedure to test the effect of the lth covariate ($\mathbf{z}_{l,i}$) on the covariance matrix, i.e.

H_0: $\gamma_{l,1,1} = \gamma_{l,1,2} = \cdots = \gamma_{l,1,K} = \gamma_{l,2,1} = \cdots = \gamma_{l,M,K} = 0$

versus

H_1: At least one $\gamma_{l,r,j} \neq \mathbf{0}$ ($j = 1,\ldots,K$ and $r = 1,\ldots,M$).

Again, to obtain D, carry out the CEM-Co algorithm with (the alternative model) and without (the null model) the lth covariate and compare the two likelihood functions. The number of degrees of freedom, in this case, is also KM (there are KM parameters $\gamma_{l,1,1}, \gamma_{l,1,2},\ldots,\gamma_{l,1,K}, \gamma_{l,2,1},\ldots,\gamma_{l,M,K}$).

One condition of using the chi-squared distribution in the LRT is that the likelihood function must follow the regularity conditions.[17] The proof for these conditions is straightforward for the tests described in the previous paragraphs, including the fact that, under the null hypothesis, the parameters $\beta_{i,l}$ and $\gamma_{l,M,K}$ cannot be in the border of the distribution support. Note that, under H_0, this condition is valid because $\beta_{i,l}$ and $\gamma_{l,M,K}$ are real numbers for any i, l, M, and K.

It is also possible to simultaneously test the effect of a covariate in the mean and the covariance structure using the LRT similarly by adjusting the degrees of freedom.

2.3. *Estimation of the number of clusters*

As described in Algorithm 1, CEM-Co requires the number of clusters (K) as input. However, in empirical data analysis, we rarely know the number of clusters *a priori*; thus, we must estimate it. We propose to use the BIC to estimate the number of clusters. Let $\hat{L}_K$ be the maximized value of the likelihood function (Eq. (4)) of the clustering structure obtained by CEM-Co with K clusters, N be the number of items, and $R = K(1 + (3 + P)M + M^2)$ be the number of parameters estimated by the model. Then, we define the BIC for the clustering structure obtained by CEM-Co with K clusters as follows:

$$\mathrm{BIC}_K = \ln(N)R - 2\ln\left(\hat{L}_K\right). \tag{6}$$

The estimated number of clusters $\widehat{K}$ is the one which minimizes the BIC statistic (Eq. (6)). Note that, the use of BIC is only a heuristic because the model selection properties of BIC for singular models are not prominent.[13]

2.4. *CEM-Co with nonlinear effect*

Sometimes the covariates are not linearly associated with the clusters' centroids (e.g. quadratic or sigmoid relationships). In this case, the method presented in Sec. 2.1 may not be suitable. To model nonlinear relationships, we extended CEM-Co as follows.

Let $f_{j,l}$ be a function representing the strength of the lth covariate effect on the jth cluster centroid. Then, we can write the "observed" clusters centroids as follows:

$$\mu_{i,j} = \mu_j^* + f_{j,1}(\mathbf{z}_{1,i}) + f_{j,2}(\mathbf{z}_{2,i}) + \cdots + f_{j,P}(\mathbf{z}_{P,i}). \tag{7}$$

To model $f_{j,l}$, we use B-spline.[10] Let S be the number of knots and B be the polynomial degree of the B-spline. Then, we represent $f_{j,l}$ by a matrix $\mathbf{W}_{j,l}$ (with dimensions $(N \times (S + B))$), where the ith row represents the spline basis of the lth covariate effect on the ith item in the jth cluster. We define the matrix $\mathbf{W}_j$ $(N \times (P\ (S + B)))$ as the combination of $\mathbf{W}_{j,1}$, $\mathbf{W}_{j,2}$,..., $\mathbf{W}_{j,P}$ by columns. Note that, we model the nonlinear effect only for the centroids and not the covariance matrices.

To obtain the CEM-Co algorithm with nonlinear covariates effects, replace $\mathbf{z}_l$ by the B-spline basis represented by $\mathbf{W}_{j,l}$ in Algorithm 1. To statistically test the effect of a single covariate, consider the number of degrees of freedom as KM (S +B) in the LRT (Sec. 2.2). To estimate the number of clusters (Sec. 2.3), consider the number of parameters as $K(1 + M\ (3 + P\ (S + B)) + M^2)$.

3. Simulations

To evaluate the performance of CEM-Co in clustering the items, the power of the LRT, and the accuracy of the BIC in estimating the number of clusters, we designed three scenarios.

Scenario 1. Covariates with linear effects on the clusters' centroids.

(1) Set the number of clusters ($K = 2$), the number of dimensions ($M = 5$), and the number of covariates ($P = 5$).
(2) Set the centroids (means) of the $K = 2$ clusters to $\mu^*_1 = (0, 0, 0, 0, 0)$ and $\mu^*_2 = (0.2, 0.2, 0.2, 0.2, 0.2)$.
(3) Set the effects of the covariates on each cluster (β) as the same as estimated by CEM-Co in the actual dataset (Sec. 4).
(4) For each item ($i = 1,\ldots,N$), simulate $P = 5$ covariates, two from a normal distribution with zero mean and unit variance ($\mathbf{z}_{1,i} \sim N\ (0, 1)$, $\mathbf{z}_{2,i} \sim N\ (0, 1)$) and three from binomial distributions with parameters equal to 0 .4, 0.25, and 0.15 ($\mathbf{z}_{3,i} \sim B(0.4)$, $\mathbf{z}_{4,i} \sim B(0.25)$, $\mathbf{z}_{5,i} \sim B(0.15)$).
(5) Simulate $N/2$ items for each cluster from a multivariate normal distribution with means equal to $\mu_{i,j} = \mu^*_{\mathrm{j}} + \beta_{1,j}\mathbf{z}_{1,i} + \beta_{2,j}\mathbf{z}_{2,i} + \beta_{3,j}\mathbf{z}_{3,i} + \beta_{4,j}\mathbf{z}_{4,i} + \beta_{5,j}\mathbf{z}_{5,i}$ (for $i = 1,\ldots,N/2$ and $j = 1, 2$) and a covariance matrix equals to a (5×5) identity matrix multiplied by 0.03 for each cluster.

Scenario 2. Covariates with linear effects on the covariance matrices.

(1) Set the number of clusters ($K = 4$), the number of dimensions ($M = 2$), and the number of covariates ($P = 1$).
(2) Set the centroids (means) of the $K = 4$ clusters to $\mu^*_1 = (0, 0)$, $\mu^*_2 = (0, 1)$, $\mu^*_3 = (1, 0)$, and $\mu^*_4 = (1, 1)$.
(3) Set the effects of the covariate on each cluster to $\beta_1 = (0.3, 0.3)$, $\beta_2 = (-0.3, -0.3)$, $\beta_3 = (0.3, -0.3)$, and $\beta_4 = (-0.3, 0.3)$.
(4) Simulate one covariate for each item ($i = 1,\ldots,N$) from a normal distribution with unit mean and unit variance ($\mathbf{z}_{1,i} \sim N(1, 1)$).
(5) Let $\mathbf{E}_j$ be a (2×2) matrix with 0.1 at the diagonal, $\sigma_{r,j} = 1$ (for $r = 1, 2$ and $j = 1, 2, 3, 4$), $\gamma_{1,r,j} = w_j$ (for $r = 1, 2$, $w_1 = 1$, $w_2 = 1$, $w_3 = 1$, and $w_4 = 10$), and $\mathbf{L}_{i,j}$ be a (2×2) matrix as defined in Sec. 2.1. Then, simulate N/K items from a multivariate normal distribution with mean and covariance matrix equal to $\mu_{i,j} = \mu^*_j + \beta_j\mathbf{z}_{1,i}$ and $\boldsymbol{\Sigma}_{i,j} = \mathbf{L}_{i,j}\mathbf{E}_j\mathbf{L}_{i,j}$, respectively, for each cluster $j = 1,\ldots,K$.

Scenario 3. Covariates with nonlinear effects on the clustering structures. To simulate nonlinear covariate effects on the clusters' centroids, replace step (5) of scenario 1 by

- Let $f_{j,1}\left(\mathbf{z}_{1,i}\right) = \beta_j \mathbf{z}_{1,i} + \beta_j \mathbf{z}_{1,i}^2$ be a quadratic function representing the strength of the covariate effect on the jth cluster centroid. Then, simulate N/K items from a multivariate normal distribution with mean equals to $\mu_{i,j} = \mu_j^* + f_{j,1}(\mathbf{z}_{1,i})$, and a covariance matrix as a (2×2) identity matrix multiplied by 0.1, for each cluster $j = 1,\ldots,K$.
- For scenario 1, the number of items varied in $N = 120$, 240, and 360. Note that, we designed this scenario to be similar to the actual data described in Sec. 4. For scenarios 2 and 3, the number of items varied in $N = 200$, 300, 400, and 800. For scenario 3, we modeled $f_{j,1}$ with a cubic B-spline with four degrees of freedom. We carried out 300 repetitions for each scenario and each number of items (N).

We also designed a fourth scenario to explore the effects of applying a dimensionality reduction technique such as principal component analysis (PCA).

Scenario 4. Dimensionality reduction effect analysis.

(1) Set the number of clusters ($K = 4$), the number of dimensions ($M = 45$), the number of covariates ($P = 1$), and the number of item ($N = 400$).
(2) Set the centroids (means) of the $K = 4$ clusters to $\mu_1^* = (1,\ldots,1)$, $\mu_2^* = (0.3,\ldots,0.3)$, $\mu_3^* = (1,\ldots,1)$, and $\mu_4^* = (0.3,\ldots,0.3)$ (each vector has size M).
(3) Set the effects of the covariate on each cluster to $\beta_1 = (0.3,\ldots,0.3)$, $\beta_2 = (0.3, 0.3)$, $\beta_3 = (0.15,\ldots,0.15)$, and $\beta_4 = (0.15, 0.15)$ (each vector has size M).
(4) Simulate one covariate for each item ($i = 1,\ldots,N$) from a normal distribution with zero mean and unit variance ($\mathbf{z}_{1,i} \sim N\,(0, 1)$).
(5) Simulate N/K items for each cluster from a multivariate normal distribution with means equal to $\mu_{i,j} = \mu_j^* + \beta_{1,j}\mathbf{z}_{1,i}$ (for $i = 1,\ldots,N/K$ and $j = 1, 2, 3, 4$) and a covariance matrix (45×45) equals to the covariance matrix estimated in the stage I NSCLC dataset described in Sec. 4.

Then, we reduced the dimensionality to M' using PCA. We varied the number of principal components (PCs) in $M' = 1, 2,\ldots,5$.

4. Stage I Lung Adenocarcinoma

4.1. *Stage I NSCLC dataset*

We collected a set of clinically annotated gene expression data composed of 45 genes (*ADAM10*, *ADAM19*, *ADAM8*, *ALOX15B*, *ATF2*, *CENPF*, *CXCL1*, *ETS2*, *FOSL1*, *GAD1*, *GAPDH*, *GNG4*, *GNG11*, *GRB10*, *HMGA2*, *HOPX*, *HSPA8*, *ID1*, *IGFBP3*, *IGFBP6*, *IL1A*, *IL1RN*, *ITGB8*, *ITPR1*, *MMP12*, *MTHFD2*, *MVK*, *NDRG1*, *PAK2*, *PHLDA2*, *PIK3CD*, *RAC2*, *S100A2*, *SEPW1*, *SERPINB5*, *SMOX*, *SPDEF*, *SPRY4*, *THRA*, *TMSB10*, *UBE2C*, *UST*, *VCP*, *VEGFA* and *YWHAQ*) and 129 stage I NSCLC patients by real-time RT-PCR with Custom TaqMan Low

Table 1. Clinicopathological characteristics of patients and their tumors. The data are numbers (%) unless otherwise stated. SD: standard deviation.

Variable	Status	n
	Mean (SD)	66.26 (10.04)
Age at diagnosis	≤65	60 (46.51)
(years)	>65	69 (53.49)
Gender	Male	59 (45.74)
	Female	74 (54.26)
Therapy	Yes	18 (13.95)
	No	110 (85.27)
	Unknown	1 (0.77)
Differentiation	1	40 (31.00)
	2	51 (39.53)
	3	38 (29.46)
Dead	Yes	48 (37.21)

Density Arrays (384-well micro fluidic cards) (Applied Biosystems). We carried out all procedures according to the manufacturer's protocol. We normalized the gene expression levels relative to the internal housekeeping control 18S gene using the method described in Ref. 25.

Table 1 summarizes clinical and pathological characteristics. The Institutional Review Board of the University of Michigan approved this study.

4.2. *Gene expression pre-processing*

We re-scaled gene expression data for zero mean and unit variance. Scaling the gene expression data for unit variance is crucial. Otherwise, a gene expression with high variance will have a higher weight than a variable with low variance. Then, we applied a PCA for dimension reduction and selected (by the elbow method) five PCs associated with the five largest eigenvalues. Altogether they represent 51.87% of the variance. These five PCs represent new dimensions of stage I NSCLC patients. Thus, we have $N = 129$ individuals, $M = 5$ dimensions, and $P = 4$ covariates (age at diagnosis, gender, therapy, and differentiation).

5. Results and Discussions

5.1. *Simulations*

To evaluate the performance of CEM-Co, we compared it against three other usual approaches, namely

(1) CEM: applying the standard classification EM algorithm on the original dataset, i.e. without taking into account the covariates;
(2) CEM-dimension: applying the standard CEM algorithm but considering the covariates as additional dimensions of the items; and
(3) CEM-partial: applying the standard CEM algorithm on the residues of the linear regression model, where the response variable is the item, and the explanatory variables are the covariates.

The three-step approach outputs the same clustering structure of CEM (the three-step method clusters the data using CEM and then identifies the association between the items and covariates). Therefore, we did not compare CEM-Co with the three-step approach.

First, we analyzed the case the covariates are linearly associated with the centroids. We calculated each repetition's adjusted rand index to evaluate how well the algorithms cluster the items. We carried out all four algorithms on the same set of items. Then, we calculated the differences between the adjusted rand indices obtained by CEM-Co minus the one obtained by each of the three alternative methods (Fig. 2(a)). The greater this difference, the better CEM-Co is than the alternative approach. Figure 2(a) shows that the performance of CEM-Co is better than all the alternative approaches. Figure 2(b) shows the frequency with the BIC selected for each number of clusters. As expected, the BIC accurately estimated the number of clusters ($K = 2$) for all examined sample sizes. Figure 2(c) shows the receiver operating characteristic (ROC) curves obtained by testing the covariate effect. To verify the type I error control, we simulated a dataset described in Sec. 3 — scenario 1 but considering no covariate effect (under the null hypothesis). When $\beta = 0$ (no covariate effect), the ROC curve (solid line) should be at the diagonal. Indeed, for $N = 360$, the LRT effectively controlled the type I error (the solid line is at the diagonal). However, for $N = 120, 240$, the ROC curves (solid lines) are above the diagonal, i.e. it rejected the null hypothesis more than expected by the p-value threshold. Then, instead of using the LRT, we propose a parametric bootstrap procedure for small N. For $N = 120$ and p-value thresholds at $0.01, 0.05$, and 0.10, the empirical false-positive rates for the bootstrap-based test (with 1000 bootstrap samples) were $0.01, 0.07$, and 0.16, respectively. In contrast, the LRT was $0.06, 0.17$, and 0.27. Thus, for small N, the bootstrap procedure better controls the type I error than the LRT. Under the alternative hypothesis, the proportion of rejected null hypothesis (power of the test) by the LRT increases as the number of items (N) and/or the covariate effect increases (dashed lines).

We also analyzed the case the covariate affects the covariance. Figure 3(a) shows that the performance of CEM-Co is better than all the alternative approaches. Figure 3(b) indicates that the BIC accurately estimated the correct number of clusters ($K = 4$) as the sample size increased. Figure 3(c) shows that, under the

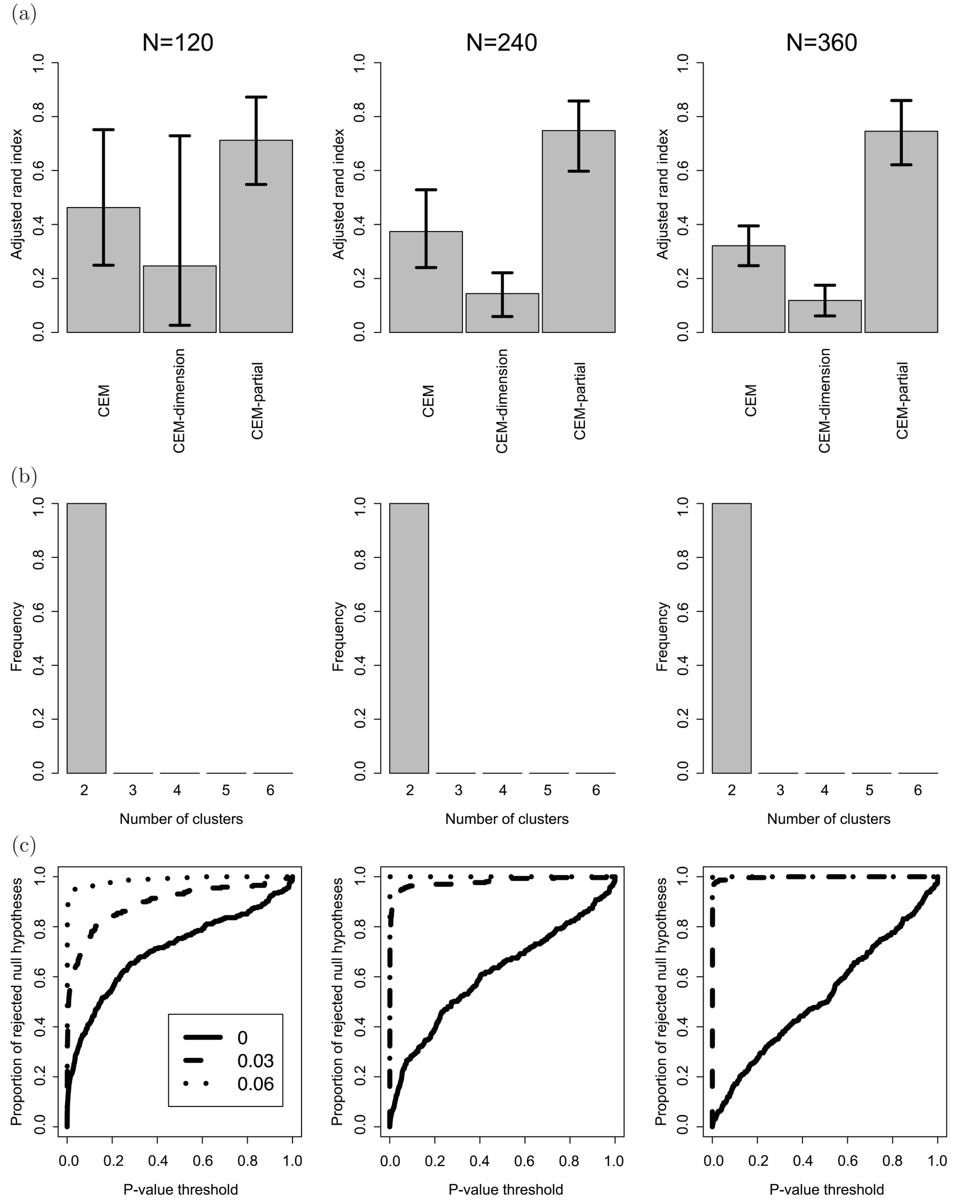

Fig. 2. Linear covariate effect on the centroids. (a) Differences between the adjusted rand indices obtained by CEM-Co and the alternative methods (CEM, CEM-dimension, CEM-partial, and one-step). The greater this difference, the better CEM-Co performs than the alternative method. Error bars represent the 90% confidence intervals. (b) Estimation of the number of clusters. The bars represent the frequency BIC selected for the indicated number of clusters. For all evaluated numbers of items (N), BIC correctly selected the number of clusters as two. (c) ROC curves. The area below the curve represents the power of the statistical test. The solid line represents the covariate effect under the null hypothesis ($\beta = 0$). Dashed lines represent the test under the alternative hypothesis ($\beta > 0$). The LRT power increases following the number of items (N) and covariate effect strength. For a small number of items ($N = 120,\ 240$), the LRT did not control the type I error (the solid line is above the diagonal).

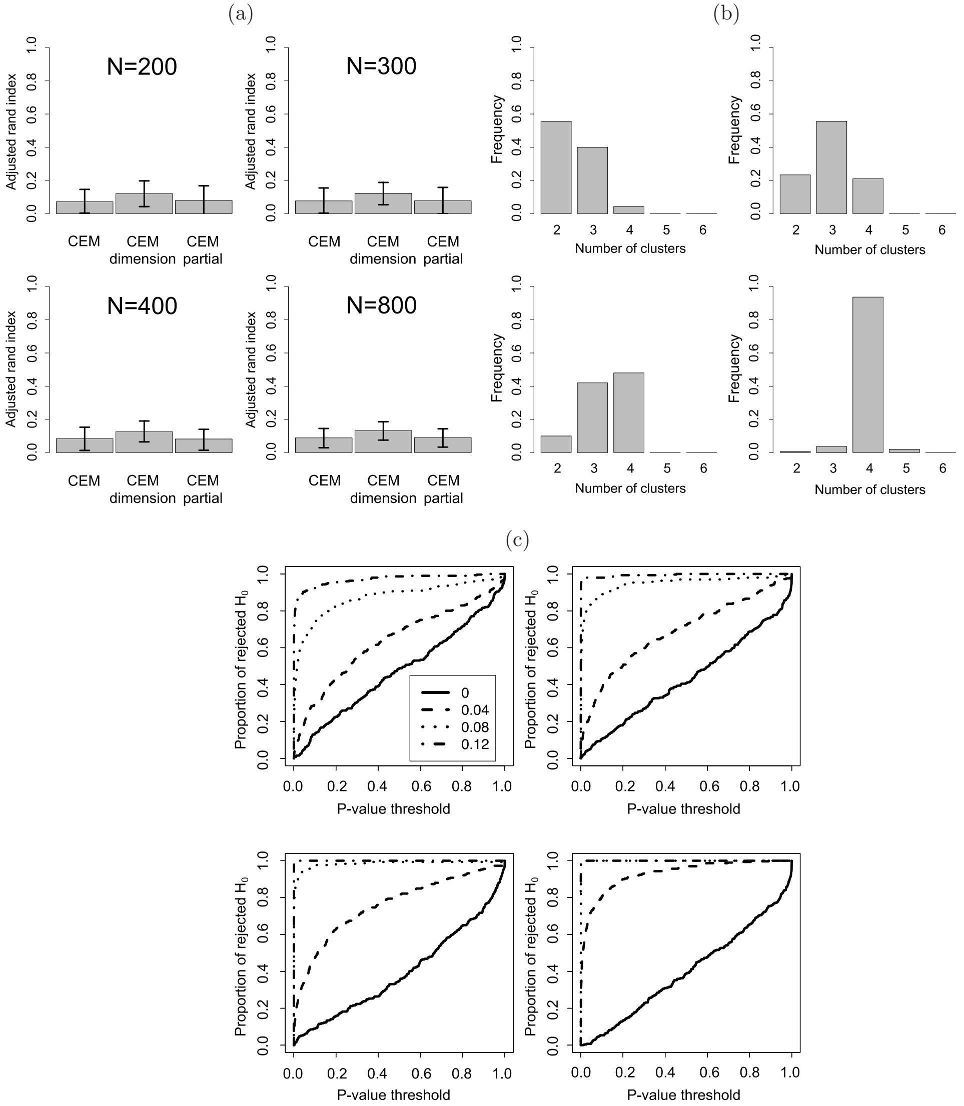

Fig. 3. Linear covariate effect on the covariances. (a) Differences between the adjusted rand indices obtained by CEM-Co and the alternative methods (CEM, CEM-dimension, CEM-partial, and one-step). The greater this difference, the better CEM-Co performs than the alternative methods. Error bars represent the 90% confidence intervals. (b) Estimation of the number of clusters. The bars represent the frequency BIC selected for the indicated number of clusters. As the number of items increases (N), BIC converges to the correct number of clusters ($K = 4$). (c) ROC-like curves. The area below the curve represents the power of the statistical test. The solid line represents the covariate effect under the null hypothesis ($\beta = 0$). Dashed lines represent the test under the alternative hypothesis ($\beta > 0$). The power of the LRT increases proportionally to the number of items (N) and covariate effect strength.

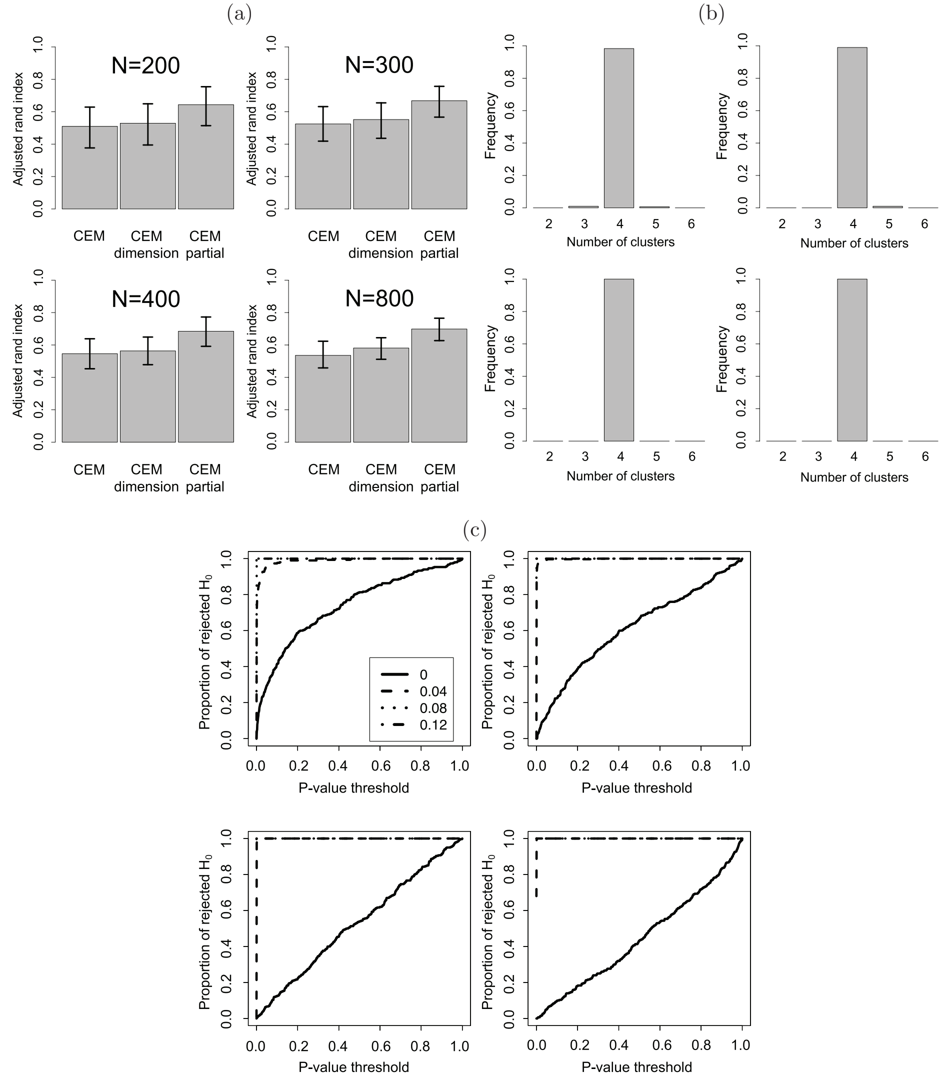

Fig. 4. Nonlinear covariate effect on the centroids. (a) Differences between the adjusted rand indices obtained by CEM-Co and the alternative methods (CEM, CEM-dimension, and CEM-partial). The greater this difference, the better CEM-Co performs than the alternative methods. Error bars represent the 90% confidence intervals. (b) Estimation of the number of clusters. The bars represent the frequency BIC selected for the indicated number of clusters. For all evaluated numbers of items (N), BIC correctly selected the number of clusters as four. (c) ROC-like curves. The area below the curve represents the power of the statistical test. The solid line represents the covariate effect under the null hypothesis ($\beta = 0$). Dashed lines represent the test under the alternative hypothesis ($\beta > 0$). The power of the LRT increases proportionally to the number of items (N) or covariate effect strength.

null hypothesis, the LRT effectively controlled the type I error (the solid line is at the diagonal). Under the alternative hypothesis, the proportion of rejected null hypothesis (power of the test) by the LRT increases as the number of items (N) and/or the covariate effect increases (dashed lines).

Besides, we analyzed the case covariates that are not linearly associated with the items. Figure 4(a) shows that the performance of CEM-Co is better than all the alternative approaches. Figure 4(b) indicates that the BIC accurately estimated the correct number of clusters ($K = 4$) for all evaluated sample sizes. Figure 4(c) shows that, under the null hypothesis, the LRT effectively controlled the type I error (the solid line is at the diagonal) as the sample size increased. Under the alternative hypothesis, the proportion of rejected null hypothesis (power of the test) by the LRT increases as the number of items (N) and/or the covariate effect increases (dashed lines).

Finally, we explored the effects of applying a dimensionality reduction technique, such as PCA. Figure 5 shows several PCs that optimize the adjusted rand index. This behavior is under what was described by Ref. 12. Ding and He[12] proved that the first $K - 1$ PCs span the cluster centroid subspace. In our example, the number of clusters is $K = 4$. Therefore, the optimal quantity of PCs is three.

5.2. *MNIST dataset*

To assess the performance of the CEM-Co algorithm, we applied our proposal to the well-known MNIST dataset. The MNIST dataset comprises approximately 70,000 hand-written digit images ranging from 0 to 9. Since this dataset does not present covariates, such as the gender or age of the digit writers, the application of CEM-Co resumes to the traditional CEM. In other words, CEM is a particular case of CEM-Co.

Due to computational constraints, we randomly sampled 10,000 digits from approximately 70,000 images. By comparing the clustering results with the known labels, we obtained an adjusted rand index of 0.38. This result is better than the performance of similar clustering algorithms, such as fuzzy c-means, on this dataset.[16]

5.3. *Stage I lung adenocarcinoma*

Finally, we applied CEM-Co and the other three alternative methods on stage I NSCLC gene expression data. Lung cancer is the most common malignancy world-wide, with an estimated 228,820 new cases and 135,720 deaths in 2020 only in the United States (American Cancer Society). There are two main types of lung cancer: NSCLC and small cell lung cancer (SCLC). They are approximately 84% and 13% of lung cancers, respectively (American Cancer Society).

Among the stage I NSCLC patients, it is estimated that 30–40% relapse and 10–30% die due to recurrence.[32] This data suggests at least one subgroup that could

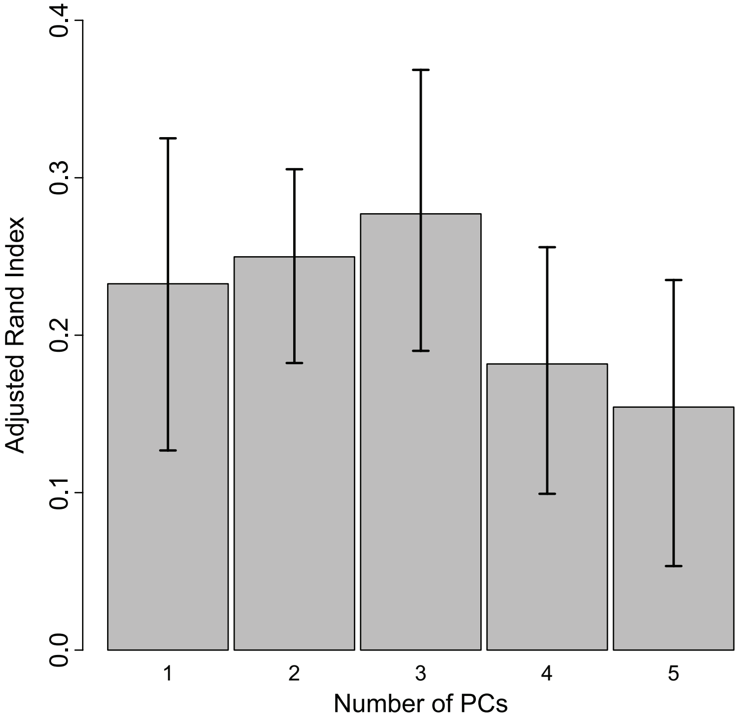

Fig. 5. Adjusted rand index by the number of PCs used at the simulations of scenario 4. The greater the adjusted rand index, the better the approach separating the actual clusters. Error bars represent the 90% confidence intervals. Note that, the actual number of clusters is $K = 4$. Therefore, the optimum number of PCs should be three.

benefit from additional therapy, whether appropriately identified. Thus, further stratification and consequent identification of these individuals become necessary. We hypothesize that one of the reasons current attempts to stratify stage I NSCLC fail is covariate effects. For example, we know that the age at diagnosis (and many others) is associated with this disease.

We only analyzed the case that the effect of covariates may be linearly associated with the centroids. We did not investigate covariates' impact on the covariances nor the nonlinear case because the number of parameters becomes higher than the number of items/individuals. We considered all the available covariates: the age at diagnosis, gender, adjuvant therapy, and differentiation.

We used the BIC described in Sec. 2.3 to estimate the number of clusters. The estimated amount of groups was two for all clustering algorithms (Fig. 6).

We verified whether the obtained two clusters presented different phenotypes after clustering the individuals using CEM-Co, CEM, CEM-dimension, and CEM-partial. One clinically relevant outcome of cancer is survival time. Then, we analyzed the association of these two clusters with the survival outcome using a Kaplan–Meier (KM) non-parametric curve (Fig. 7). Also, we used a multivariate

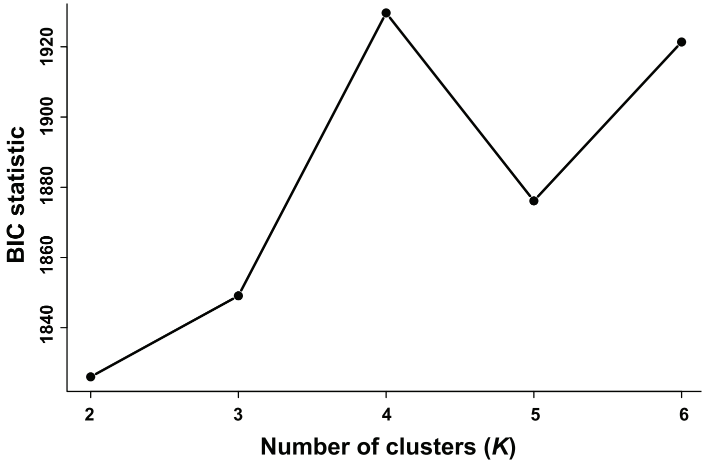

Fig. 6. Estimating the number of clusters in stage I NSCLC gene expression data using the BIC for CEM-Co. The BIC statistic reaches its minimum value for $\widehat{K} = 2$. Therefore, the estimated number of clusters in this dataset is $\widehat{K} = 2$.

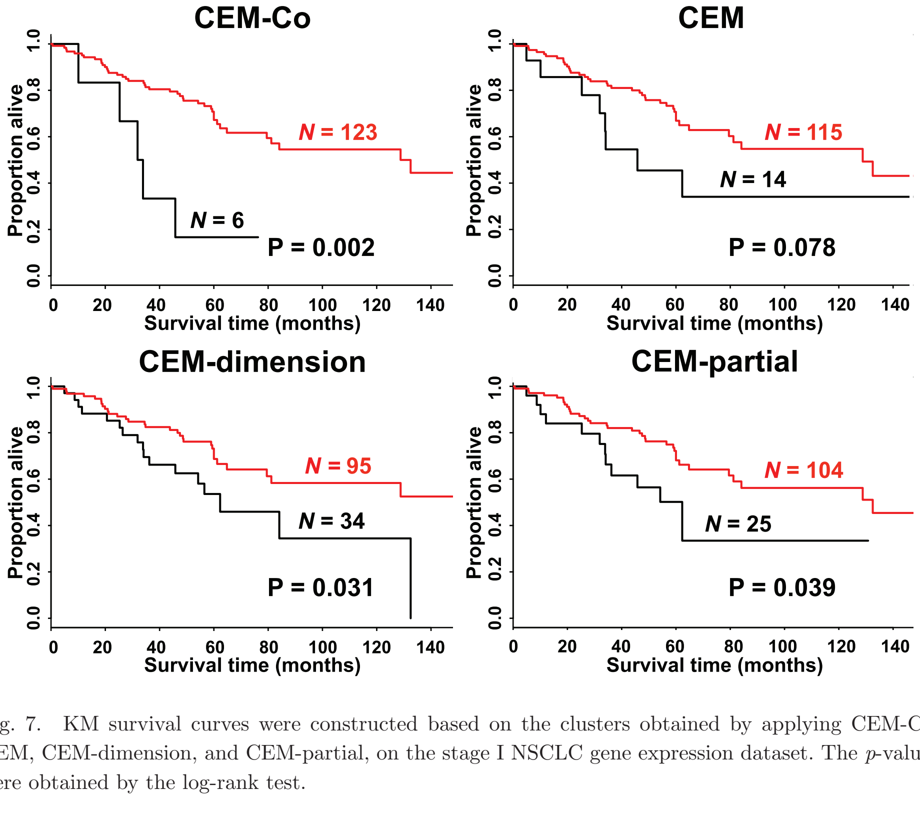

Fig. 7. KM survival curves were constructed based on the clusters obtained by applying CEM-Co, CEM, CEM-dimension, and CEM-partial, on the stage I NSCLC gene expression dataset. The p-values were obtained by the log-rank test.

Cox proportional hazards model with age at diagnosis, gender, adjuvant therapy, and differentiation as covariates (Table 2).

By setting a p-value threshold at 5% in the KM analyses, only the CEM algorithm did not stratify the stage I NSCLC dataset into two clusters with different survival outcomes (Fig. 7). However, notice that, the KM analysis does not remove the effects of the covariates. Thus, we also analyzed using the Cox regression model, which considers the covariates. In this case, only CEM-Co identified a subgroup of stage I NSCLC with a statistically poorer survival outcome ($p < 0.001$) (Table 2).

We also analyzed the same dataset using the UMAP dimensionality reduction method[21] and the HDBSCAN clustering algorithm.[5] UMAP reduced the dimensionality to three. Then, we clustered the individuals using the HDBSCAN method

Table 2. Multivariate survival analyses were performed using the Cox regression model.

Method	Variables	Hazard ratio (95%CI)	p-value
CEM-Co	Cluster (red/black)	0.117 (0.041–0.334)	**<0.001**
	Age at diagnosis	1.084 (1.048–1.121)	**<0.001**
	Gender (male/female)	1.231 (0.668–2.268)	0.5045
	Therapy (yes/no)	2.545 (1.223–5.296)	**0.013**
	Differentiation (2/1)	1.348 (0.605–3.003)	0.466
	Differentiation (3/1)	2.015 (0.853–4.762)	0.1101
CEM	Cluster (red/black)	0.436 (0.181–1.052)	0.0646
	Age at diagnosis	1.082 (1.045–1.120)	**<0.001**
	Gender (male/female)	1.220 (0.667–2.232)	0.519
	Therapy (yes/no)	2.151 (1.049–4.410)	**0.037**
	Differentiation (2/1)	2.064 (0.701–3.438)	0.278
	Differentiation (3/1)	2.064 (0.861–4.949)	0.104
CEM-dimension	Cluster (red/black)	0.599 (0.251–1.429)	0.248
	Age at diagnosis	1.077 (1.041–1.115)	**<0.001**
	Gender (male/female)	0.201 (0.671–2.228)	0.511
	Therapy (yes/no)	1.378 (0.491–3.871)	0.543
	Differentiation (2/1)	1.645 (0.722–3.750)	0.236
	Differentiation (3/1)	2.545 (1.108–5.846)	**0.028**
CEM-partial	Cluster (red/black)	0.574 (0.290–1.137)	0.111
	Age at diagnosis	1.076 (1.040–1.114)	**<0.001**
	Gender (male/female)	0.796 (0.691–2.284)	0.459
	Therapy (yes/no)	2.013 (0.975–4.156)	0.058
	Differentiation (2/1)	1.651 (0.737–3.696)	0.223
	Differentiation (3/1)	2.538 (1.102–5.841)	**0.029**

Notes: CI: confidence interval. p-values < 0.05 are in bold.

with hyperparameter optimization. We set the minimum cluster size to six, i.e. the smallest cluster size obtained using CEM-Co. As a result, we got five clusters. However, by analyzing these five clusters using the Cox regression model, we could not identify any cluster associated with survival (all p-values were greater than 0.2). We also varied the number of dimensions obtained by UMAP to two and four. We varied the smallest cluster size to ten and twenty. We got the same results. Thus, our results obtained using UMAP and HDBSCAN are robust to the selected number of dimensions and smallest cluster size parameter.

The reason why only CEM-Co was able to identify the poorer subgroup can be explained by how the covariates affect the clustering structure. First, except by therapy ($p = 0.219$), all other covariates significantly affect the clusters' centroids: age at diagnosis ($p = 0.002$), gender ($p = 0.001$), and differentiation ($p = 0.007$). This result confirms our hypothesis that covariates affect stage I NSCLC stratification. Second, the covariates affect the clusters' centroids and items' dimensions differently (Fig. 8). However, both CEM-dimension and CEM-partial assume that covariates affect the cluster centroids equally.

To verify the robustness of our results, we repeated all analyses by considering the number of clusters as $K = 3$ (the second-best choice by BIC (Fig. 6)). In this case, the largest group (red curve in Fig. 7) was divided into two clusters (by all clustering methods). The Cox regression model for the clustering structure obtained by CEM-Co indicated that the smallest group presented a poorer survival outcome than the other two clusters. The two largest clusters did not show different statistical survival outcomes between them. The three clusters did not present statistically different survival outcomes for the groups obtained by CEM, CEM-dimension, and CEM-partial. Therefore, the conclusions derived for $K = 3$ were the same at $K = 2$.

Instead of selecting $M = 5$ PCs, we reanalyzed data considering one, four, and six PCs. The conclusions are the same for one ($M = 1$, representing 23.49% of the variance) and four PCs ($M = 4$, representing 46.41% of the variance) at $M = 5$. Whether we consider six PCs ($M = 6$, representing 55.8% of the variance), CEM-Co clustered the dataset into three groups. In other words, CEM-Co split the largest group into two groups. The smallest cluster presented a poorer survival outcome than the other two groups. We did not observe any statistical difference between the two largest clusters. Therefore, the results seem robust even for different numbers of PCs.

Here, we illustrated the importance of CEM-Co in stratifying a stage I NSCLC dataset by minimizing the effects of age at diagnosis, gender, adjuvant therapy, and differentiation. However, we imagine that the flexibility of CEM-Co would make it applicable to other areas where clustering is a source of concern. For example, in large neuroimaging projects, it is known that the site where the data is collected strongly affects the results.

The EM algorithm has interesting theoretical properties: monotonicity and convergence to a stationary value. The proofs for our proposal are similar to the one

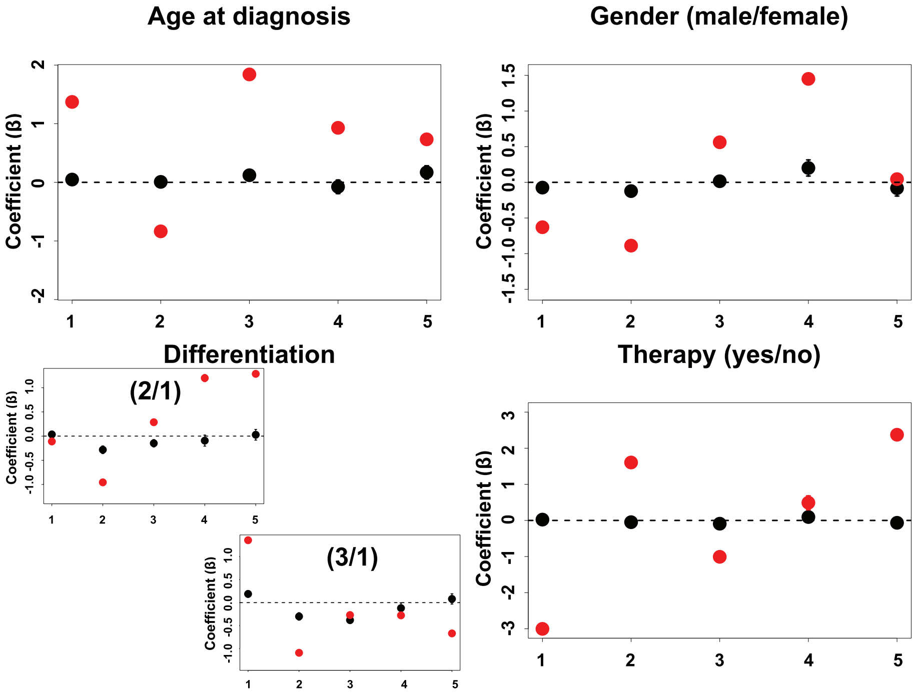

Fig. 8. Covariates effects estimated by CEM-Co on stage I NSCLC. The x- and y-axes represent the M = 5 dimensions of the items (PCs) and the coefficients (β) for each covariate effect. Red and black dots represent the patients with good and poor prognoses, respectively. Error bars represent the 95% confidence intervals (some cannot be seen because they are too small). Except by therapy ($p = 0.219$), all other covariates are significantly associated with the clusters' centroids (age at diagnosis – $p = 0.002$, gender – $p = 0.001$, and differentiation – $p = 0.007$). Note that, the coefficients (covariates effects) differ between clusters (between red and black dots) and the five dimensions. The small cluster size can probably explain the coefficients close to zero for the cluster with a worse prognosis (black dots).

presented in Ref. 22. One just needs to condition on the observed value of $\mathbf{z}_i = (\mathbf{z}_{1,i}, \mathbf{z}_{2,i}, \ldots, \mathbf{z}_{P,i})$.

One disadvantage of CEM-Co in the present form is that it only can be applied to data that a mixture of normal distributions can model. For example, we cannot apply CEM-Co to datasets represented by discrete variables. However, we can obtain a similar method for discrete data using an appropriate probability distribution.

Acknowledgments

We thank Alexandre Galvão Patriota and Daniel Yasumasa Takahashi for reading the manuscript and providing valuable critiques and comments. C.E.M.R. was partially supported by the São Paulo Research Foundation (FAPESP

2018/21934-5). A.N. was partially supported by CAPES, MEXT (15K18426), and the JSPS for the "Institutional Program for Young Researcher Overseas Visits". G.C. and D.G.B. were partially supported by NCI (P30CA046592). N.G. was partially supported by JST (12103436). A.F. was partially supported by FAPESP (2018/17996-5 and 2018/21934-5), CNPq (306811/2022-7), CAPES (Finance code 001), Alexander von Humboldt Foundation, The Academy of Medical Sciences — Newton Fund and Wellcome Leap.

Appendix A

Like the usual CEM algorithm, the CEM-Co algorithm has some interesting properties. Among these properties, we highlight the monotonicity and convergence to a stationary value of the algorithm.

Theorem A.1. *The CEM-Co optimization algorithm is monotonic between iterations.*

Proof. Monotonicity means that $L(\Theta^{k+1}) \geq L(\Theta^{k})$, for k = 0,1,2,..., then after each new iteration of the steps expectation and maximization, the estimated likelihood is higher. Let $g_c\left(\mathbf{y}|\mathbf{z}_1,\ldots,\mathbf{z}_N\right);\theta\big)$ be the probability density distribution of the random vector $\mathbf{y}$ representing the complete data given the covariates effects $(\mathbf{z})$. Instead of looking at the complete data $\mathbf{y}$, we observe incomplete data $\mathbf{x}$, where $\mathbf{x}=\mathbf{x}(\mathbf{y})$. The complete data $(\mathbf{y})$ represent the situation in which we know the real cluster of each observation. The complete log-likelihood function can be written as follows:

$$\begin{aligned}&\log L_c\left(\theta|\mathbf{x}_1,\ldots,\mathbf{x}_N,\mathbf{z}_1,\ldots,\mathbf{z}_N\right) = \\ &\log g_c\left(\mathbf{y}|\mathbf{x}_1,\ldots\mathbf{x}_N,\mathbf{z}_1,\ldots,\mathbf{z}_N\right);\theta\big).\end{aligned} \tag{A.1}$$

In general, we can describe the steps of *expectation* and *maximization* as follows: E-step: Calculate

$$Q\left(\theta;\theta^{k}\right) = E_{\theta^{k}}\left(\log L_c\left(\theta|\mathbf{x}_1,\ldots,\mathbf{x}_N\right)|\mathbf{z}\right), \tag{A.2}$$

M-step: Choose θ^{k+1} that maximizes $Q(\theta;\theta^{k})$ for all $\theta^{k+1} \in \mathbf{\Omega}$.

Furthermore, these two steps are alternated until the algorithm converges.

We will show that the likelihood function of incomplete data $L(\theta)$ increases monotonically, i.e.

$$L\left(\theta^{k+1}|\mathbf{x}_1,\ldots,\mathbf{x}_N\right) \geq L\left(\theta^{k}|\mathbf{x}_1,\ldots,\mathbf{x}_N\right), \tag{A.3}$$

for k = 1,2, ...

Let

$$h\left(\mathbf{y}|\mathbf{x},\mathbf{z}_1,\ldots,\mathbf{z}_N;\theta\right) = \frac{g_c\left(\mathbf{y}|\mathbf{z}_1,\ldots,\mathbf{z}_N;\theta\right)}{g\left(\mathbf{x}|\mathbf{z}_1,\ldots,\mathbf{z}_N;\theta\right)}, \tag{A.4}$$

be the conditioned density of **Y** given **X**, where **Y** represents the complete data and X represents the incomplete data. Then, we can rewrite the likelihood function (Eq. (A.1)) as follows:

$$\begin{aligned} L\left(\theta|\mathbf{x}_1,\ldots,\mathbf{x}_N,\mathbf{z}_1,\ldots,\mathbf{z}_N\right) &= \log g\left(\mathbf{x}|\mathbf{z}_1,\ldots,\mathbf{z}_N;\theta\right) = \\ \log g_c\left(\mathbf{y}|\mathbf{z}_1,\ldots,\mathbf{z}_N;\theta\right) &- \log h\left(\mathbf{y}|\mathbf{x},\mathbf{z}_1,\ldots,\mathbf{z}_N;\theta\right) = \\ \log L_c\left(\theta|\mathbf{x}_1,\ldots,\mathbf{x}_N,\mathbf{z}_1,\ldots,\mathbf{z}_N\right) &- \log h\left(\mathbf{y}|\mathbf{x},\mathbf{z}_1,\ldots,\mathbf{z}_N;\theta\right). \end{aligned} \tag{A.5}$$

Calculating the expected value of both sides of Eq. (A.5) conditional to the distribution of **Y** given **Z**, we obtain

$$\begin{aligned} L\left(\theta|\mathbf{x}_1,\ldots,\mathbf{x}_N,\mathbf{z}_1,\ldots,\mathbf{z}_N\right) &= \\ E_{\theta(k)}\{L_c\left(\theta|\mathbf{x}_1,\ldots,\mathbf{x}_N\right))|\mathbf{z}\} - E_{\theta(k)}\{\log h\left(\mathbf{Y}|\mathbf{x},\mathbf{z}_1,\ldots,\mathbf{z}_N\right);\theta)|\mathbf{z}\} &= \\ \mathbf{Q}\left(\theta;\theta^{(k)}\right) - \mathbf{H}\left(\theta;\theta^{(k)}\right). \end{aligned} \tag{A.6}$$

Using Eq. (A.6), we obtain

$$\begin{aligned} \log L\left(\theta^{(k+1)}|\mathbf{x}_1,\ldots,\mathbf{x}_N,\mathbf{z}_1,\ldots,\mathbf{z}_N\right) &- \\ \log L\left(\theta^{(k)}|\mathbf{x}_1,\ldots,\mathbf{x}_N,\mathbf{z}_1,\ldots,\mathbf{z}_N\right) &= \\ \left\{\mathbf{Q}\left(\theta^{(k+1)};\theta^{(k)}\right) - \mathbf{Q}\left(\theta^{(k)};\theta^{(k)}\right)\right\} &- \\ \left\{\mathbf{H}\left(\theta^{(k+1)};\theta^{(k)}\right) - \mathbf{H}\left(\theta^{(k)};\theta^{(k)}\right)\right\}&. \end{aligned} \tag{A.7}$$

Note that, $\mathbf{Q}\left(\theta^{(k+1)};\ \theta^{(k)}\right) - \mathbf{Q}\left(\theta^{(k)};\theta^{(k)}\right) \geq 0$ because it is the maximization of the complete likelihood function itself. Then, it remains to prove that

$$\mathbf{H}\left(\theta^{(k+1)};\theta^{(k)}\right) - \mathbf{H}\left(\theta^{(k)};\theta^{(k)}\right) \leq 0 \tag{A.8}$$

We have that for any q

$$\begin{aligned} \mathbf{H}\left(\theta;\theta^{(k)}\right) - \mathbf{H}\left(\theta^{(k)};\theta^{(k)}\right) &= \\ E_{\theta^{(k)}}\left[\log\left\{\frac{h\left(\mathbf{y}|\mathbf{x},\mathbf{z}_1,\ldots,\mathbf{z}_N\right);\theta)}{h\left(\mathbf{y}|\mathbf{x},\mathbf{z}_1,\ldots,\mathbf{z}_N\right);\theta^{(k)})}\right\}|\mathbf{z}\right]& \\ \leq \log\left[E_{\theta^{(k)}}\left\{\frac{h\left(\mathbf{y}|\mathbf{x},\mathbf{z}_1,\ldots,\mathbf{z}_N\right);\theta)}{h\left(\mathbf{y}|\mathbf{x},\mathbf{z}_1,\ldots,\mathbf{z}_N\right);\theta^{(k)})}\right\}|\mathbf{z}\right]& \\ = \log\int_{\chi(\mathrm{z})} H\left(\mathbf{y}|\mathbf{x},\mathbf{z}_1,\ldots,\mathbf{z}_N;\theta\right)d\mathbf{y}& \\ = 0,& \end{aligned} \tag{A.9}$$

The inequality is a consequence of both Jensen's inequality and the concavity of the logarithmic function.

Theorem A.2. *The CEM-Co optimization algorithm converges to a stationary sequence.*

Proof. Let us prove that the CEM-Co algorithm converges to a stationary sequence. In other words, that $L(\Theta^k)$ converges monotonically to some value $L^* = L(\Theta^k)$, where L^* is a stationary value in almost all applications, and that Θ^* is the value that

$$\frac{\partial L(\Theta)}{\partial \Theta} = 0.$$

Note that, we can write the likelihood function as Eq. (A.6). Differentiating both sides of this equation, we obtain

$$\frac{\partial \log L\left(\theta | \mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N\right)}{\partial \theta}$$
$$\frac{\partial \mathbf{Q}\left(\theta; \theta^k\right)}{\partial \theta} - \frac{\partial \mathbf{H}\left(\theta; \theta^k\right)}{\partial \theta}. \tag{A.10}$$

In the proof of Theorem A.1, we showed that $\mathrm{H}(\theta;\, \theta^k) \geq \mathrm{H}(\theta^k;\, \theta^k)$ for all θ. Thus

$$\left[\frac{\partial \mathbf{H}\left(\theta; \theta^k\right)}{\partial \theta}\right]_{\theta - \theta^k} = 0. \tag{A.11}$$

We can write Eq. (A.11) for some arbitrary θ_0 as follows:

$$\frac{\partial \log L\left(\theta = \theta_0 | \mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N\right)}{\partial \theta} =$$
$$\left[\frac{\partial \mathbf{Q}\left(\theta; \theta_0\right)}{\partial \theta}\right]_{\theta - \theta_0}. \tag{A.12}$$

Suppose that θ_0 is a stationary point of the likelihood function, then

$$\frac{\partial \log L\left(\theta = \theta_0 | \mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N\right)}{\partial \theta} =$$
$$\left[\frac{\partial \mathbf{Q}\left(\theta; \theta_0\right)}{\partial \theta}\right]_{\theta - \theta_0} = 0. \tag{A.13}$$

Using Wu's proposition,[31] we obtain

$$\sup_{\theta \in \Omega} \mathbf{Q}\left(\theta; \theta_0\right) \geq \mathbf{Q}\left(\theta_0; \theta_0\right), \tag{A.14}$$

to any stationary point θ_0, that is not a local maximum. The conditions of regularity[22] ensure that all limit points of any CEM-Co algorithm solution are local maximums and that $L(\theta^k)$ converges monotonically to some local maximums.

References

1. MacQueen J, Some methods for classification and analysis of multivariate observations, *Proc Fifth Berkeley Symp Mathematical Statistics and Probability*, University of California Press, pp. 281–297, 1967.
2. Kaufman L, Rousseeuw P, *Clustering by Means of Medoids*, North-Holland, 1987.
3. Ward Jr JH, Hierarchical grouping to optimize an objective function, *J Am Stat Assoc* **58**(301):236–244, 1963.
4. McLachlan GJ, Peel D, *Finite Mixture Models*, John Wiley & Sons, 2004.
5. Ester M, Kriegel HP, Sander J, Xu X, A density-based algorithm for discovering clusters in large spatial databases with noise, *Proc Second Int Conf Knowledge Discovery and Data Mining*, AAAI Press, pp. 226–231, 1996.
6. Cheng Y, Mean shift, mode seeking, and clustering, *IEEE Trans Pattern Anal Mach Intell* **17**(8):790–799, 1995.
7. Bezdek JC, Ehrlich R, Full W, FCM: The fuzzy *c*-means clustering algorithm, *Comput Geosci* **10**(2–3):191–203, 1984.
8. Von Luxburg U, A tutorial on spectral clustering, *Stat Comput* **17**(4):395–416, 2007.
9. Bandeen-Roche K, Miglioretti DL, Zeger SL, Rathouz PJ, Latent variable regression for multiple discrete outcomes, *J Am Stat Assoc* **92**(440):1375–1386, 1997.
10. Dayton CM, Macready GB, Concomitant-variable latent-class models, *J Am Stat Assoc* **83**(401):173–178, 1988.
11. Asparouhov T, Muth́en B, Auxiliary variables in mixture modeling: Three-step approaches using M*plus*, *Struct Equ Model Multidiscip J* **21**(3):329–341, 2014.
12. Kamata A, Kara Y, Patarapichayatham C, Lan P, Evaluation of analysis approaches for latent class analysis with auxiliary linear growth model, *Front Psychol* **9**:130, 2018.
13. Nylund-Gibson K, Grimm R, Quirk M, Furlong M, A latent transition mixture model using the three-step specification, *Struct Equ Model Multidiscip J* **21**(3): 439–454, 2014.
14. Vermunt JK, Latent class modeling with covariates: Two improved three-step approaches, *Polit Anal* **18**(4):450–469, 2010.
15. Gudicha DW, Vermunt JK, Mixture model clustering with covariates using adjusted three-step approaches, in *Algorithms from and for Nature and Life*, Springer, pp. 87–94, 2013.
16. Bolck A, Croon M, Hagenaars J, Estimating latent structure models with categorical variables: One-step versus three-step estimators, *Polit Anal* **12**(1):3–27, 2004.
17. Vermunt JK, Magidson J, *Latent Gold 4.0 User's Guide*, 2005.
18. Celeux G, Govaert G, A classification EM algorithm for clustering and two stochastic versions, *Comput Stat Data Anal* **14**(3):315–332, 1992.
19. Celeux G, Govaert G, Gaussian parsimonious clustering models, *Pattern Recognit* **28**(5):781–793, 1995.
20. Dereniowski D, Kubale M, Cholesky factorization of matrices in parallel and ranking of graphs, *Int Conf Parallel Processing and Applied Mathematics*, Springer, pp. 985–992, 2003.
21. Wilks SS, The large-sample distribution of the likelihood ratio for testing composite hypotheses, *Ann Math Stat* **9**(1):60–62, 1938.
22. Hogg RV, McKean J, Craig AT, *Introduction to Mathematical Statistics*, Pearson Education, 2005.

23. Drton M, Plummer M, A Bayesian information criterion for singular models, *J R Stat Soc* **79**(2):323–380, 2017.
24. De Boor C, On calculating with B-splines, *J Approx Theory* **6**(1):50–62, 1972.
25. Schmittgen TD, Livak KJ, Analyzing real-time PCR data by the comparative *CT* method, *Nat Protoc* **3**(6):1101–1108, 2008.
26. Ding C, He X, *K*-means clustering via principal component analysis, *Proc Twenty-First Int Conf Machine Learning*, Association for Computing Machinery, pp. 29, 2004.
27. Havens TC, Bezdek JC, Leckie C, Hall LO, Palaniswami M, Fuzzy c-means algorithms for very large data, *IEEE Trans Fuzzy Syst* **20**(6):1130–1146, 2012.
28. Zhu C-Q *et al.*, Prognostic and predictive gene signature for adjuvant chemotherapy in resected non-small-cell lung cancer, *J Clin Oncol* **28**(29):4417–4424, 2010.
29. McInnes L, Healy J, Saul N, Großberger L, UMAP: Uniform manifold approximation and projection, *J Open Source Softw* **3**(29):861, 2018, doi:10.21105/joss.00861.
30. Campello RJGB, Moulavi D, Sander J, Density-based clustering based on hierarchical density estimates, in Pei J, Tseng VS, Cao L, Motoda H, Xu G (eds.), *Advances in Knowledge Discovery and Data Mining*, Springer, Berlin, pp. 160–172, 2013.
31. McLachlan G, Krishnan T, *The EM Algorithm and Extensions*, John Wiley & Sons, 2007.
32. Wu CFJ, On the convergence properties of the EM algorithm, *Ann Stat* **11**(1):95– 103, 1983.

Carlos Relvas has a Bachelor's and Master's degree in Statistics and a Ph.D. in Computer Science from the Institute of Mathematics and Statistics at the University of São Paulo (IME-USP). With over 13 years of experience as a data scientist, he has worked for five years at Itaú in Research and Development, focusing on predictive modeling and Big Data, and another four years as a lead data scientist at Nubank. He is the co-founder and head of data science at the startup DataRisk.

Asuka Nakata PhD is an Assistant Professor at the Medical Institute of Bioregulation of Kyushu University in Japan. Previously, she was an Assistant Professor at Kanazawa University. She received her Ph.D. from Nara Institute of Science and Technology in 2010. Her research focuses on cellular and molecular biology in cancer stem cells.

Guoan Chen MD PhD is an Associate Professor in the School of Medicine, Southern University of Science and Technology. Dr. Chen received his medical and master's degrees from Xi'an Jiaotong University, China, in 1986 and 1992, respectively. He received his Ph.D. in oncology from Peking Union Medical College in 1999. In 2000, Dr. Chen joined Professor David G. Beer's Tumor Biology Laboratory at the University of Michigan as a Research Fellow. He was promoted to the faculty as a Research Investigator, Research Assistant Professor, and Associate Research Scientist at the University of Michigan Medical School in 2004, 2010, and 2017, respectively. Dr. Chen's research focuses on genomics, proteomics, and bioinformatics in lung cancer. He utilizes state-of-the-art molecular technologies and approaches, including gene expression microarray, SNP array, miRNA array, and RNA deep sequencing, to identify molecular changes for patient survival, early diagnosis, and the molecular mechanism in lung cancer progression. He works on micro-RNAs, long non-coding RNAs, and circRNAs for lung cancer diagnosis and prognosis and the functional, mechanistic, and therapeutic application of these molecules in lung cancer. He has published over 90 SCI research articles, such as Nature Medicine, Nature Biotechnology, Cancer Cell, JNCI, J Clinical Investigation, Nature Communications, Cancer Research, Clinical Cancer Research, J Thoracic Oncology, Genome Research, Developmental Cell, and PNAS. He has been cited over 10,000 times with h-index 51, i10-index 79.

David G. Beer PhD is Professor Emeritus at the University of Michigan Medical School. His prior laboratory focused on defining the genetic alterations and the gene profiles of lung and esophageal cancer. He received his Ph.D. from the University of Colorado in 1984. He was a research fellow at the McArdle Laboratory for Cancer Research until 1987. He was an Assistant Professor at the University of Kansas Medical School before starting at the University of Michigan in 1991.

Noriko Gotoh MD PhD is a Professor at the Cancer Research Institute, Kanazawa University in Japan. Until 2013, she was an Associate Professor at the Institute of Medical Science (IMSUT), The University of Tokyo in Japan. Her laboratory studies cancer stem cells and tumor heterogeneity, focusing on breast and lung cancer, molecular mechanisms, and how cancer stem cells are maintained through interaction with cancer stem cell niche and growth factor signaling. She received an M.D. from Kanazawa University in 1989 and a Ph.D. from The University of Tokyo in 1993. Her research training was in IMSUT, The University of Tokyo, and the New York University School of Medicine.

André Fujita PhD is an Associate Professor at the Institute of Mathematics and Statistics of the University of São Paulo. He received his Ph.D. from the University of São Paulo in 2007. He was a Special Postdoctoral Researcher at RIKEN, a JICA Fellow at Kanazawa University, an Alexander von Humboldt Fellow at Leipzig University, a Newton Advanced Fellow at University College London, a FAPESP-ERC Fellow at King's College London, and a Fulbright Fellow at Boston University. His research focuses on network statistics and heart informatics.