Wong *et al. BMC Bioinformatics* (2017) 18:150
DOI 10.1186/s12859-017-1564-5

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Two-way learning with one-way supervision for gene expression data

Monica H. T. Wong[1*], David M. Mutch[2] and Paul D. McNicholas[1]

Abstract

Background: A family of parsimonious Gaussian mixture models for the biclustering of gene expression data is introduced. Biclustering is accommodated by adopting a mixture of factor analyzers model with a binary, row-stochastic factor loadings matrix. This particular form of factor loadings matrix results in a block-diagonal covariance matrix, which is a useful property in gene expression analyses, specifically in biomarker discovery scenarios where blood can potentially act as a surrogate tissue for other less accessible tissues. Prior knowledge of the factor loadings matrix is useful in this application and is reflected in the one-way supervised nature of the algorithm. Additionally, the factor loadings matrix can be assumed to be constant across all components because of the relationship desired between the various types of tissue samples. Parameter estimates are obtained through a variant of the expectation-maximization algorithm and the best-fitting model is selected using the Bayesian information criterion. The family of models is demonstrated using simulated data and two real microarray data sets. The first real data set is from a rat study that investigated the influence of diabetes on gene expression in different tissues. The second real data set is from a human transcriptomics study that focused on blood and immune tissues. The microarray data sets illustrate the biclustering family's performance in biomarker discovery involving peripheral blood as surrogate biopsy material.

Results: The simulation studies indicate that the algorithm identifies the correct biclusters, most optimally when the number of observation clusters is known. Moreover, the biclustering algorithm identified biclusters comprised of biologically meaningful data related to insulin resistance and immune function in the rat and human real data sets, respectively.

Conclusions: Initial results using real data show that this biclustering technique provides a novel approach for biomarker discovery by enabling blood to be used as a surrogate for hard-to-obtain tissues.

Keywords: Biclustering, Biomarker discovery, Finite mixture models, Microarray gene expression, Surrogate tissue

Background

With the introduction of personalized medicine, the discovery of novel biomarkers via "omics" research plays a critical role in its advancement [1]. A biomarker is defined as "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention" [2]. The behaviour of a biomarker is expected to vary among individuals, thereby allowing treatment to be "personalized" depending on that individual's (predicted) response. The ideal diagnostic tool is minimally invasive, leading researchers to investigate the use of peripheral blood cells as surrogate biopsy material, since blood is more easily accessible. The assumption is that the molecular profile of peripheral blood reflects a global overview of the physiological events occurring in different tissues throughout the body [3].

When gene expression microarrays are used for biomarker discovery, the subset of identified genes acts as the set of biomarkers [4, 5]. Returning to the idea of peripheral blood as surrogate material, a gene that exhibits correlated expression profiles in blood and a given tissue (but not other tissues) may be a biomarker of interest. In this scenario, the genes act as the observations and the blood and tissues (the samples) act as the variables. A data point in the microarray data set is thus an intensity value

*Correspondence: monica.wong@math.mcmaster.ca
[1]Department of Mathematics and Statistics, McMaster University, L8S 4L8 Hamilton, ON, Canada
Full list of author information is available at the end of the article

corresponding to a specific gene in a sample. One popular way of identifying these subsets of correlated genes across the blood and the given tissue is via clustering techniques [6].

One-way clustering methods can be restrictive in certain applications. It is not always the case that the groups of patterns found in the observations are homogeneous across all the variables; rather, it could be the case that only a subset of the variables possesses these groupings. With gene expression data, if the samples are the variables and the researcher hypothesizes that there exists homogeneous groups of samples, this would be useful information for the algorithm to have. A popular example is the discovery of leukaemia tumour subtypes based on gene expression [7]. Consequently, biclustering techniques have been developed to address this recurring issue. Biclustering, first explored by Hartigan [8], clusters both rows and columns simultaneously and results in biclusters.

Biclustering is a useful technique when the researcher suspects biclusters of variables and observations in the data, but does not understand what properties of the variables define the biclusters. For instance, in the leukaemia tumour subtype analysis, researchers initially would not have known the classes of each tumour sample (see [7] for a discussion). Here, biclustering could help to reveal these subtypes more efficiently, as done by Kluger et al. [9] for example. However, researchers may desire that the observations within biclusters satisfy a particular relationship among the variables; the biclustering method would then be one-way supervised. This technique is particularly relevant for the blood biomarker discovery application mentioned earlier. One-way supervision is effective because the researcher specifically requires a prominent relationship between the samples of blood and the samples of the tissue of interest with respect to the expression profile of a subset of genes. Additionally, the researcher explicitly requires that the expression profiles of that same subset of genes to have no relationship between the previously mentioned samples and the rest of the samples in the data set. In this way, the resulting biclusters would contain a subset of genes that is strictly correlated within blood and the tissue of interest only.

Model-based clustering

Cluster analysis identifies homogeneous groups that are relevant within a population. It is an unsupervised technique because it does not utilize existing labels to find the best homogeneous groups among a set of observations, which reflects common real-life scenarios because observations are not usually accompanied by hints about their true groupings with respect to the variables. Some popular clustering techniques include methods such as hierarchical clustering [10], k-means clustering [11], and model-based clustering (see [12] for an in-depth discussion).

In model-based clustering, group membership is estimated using a parametric finite mixture model, which can be denoted

$$f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g f_x\left(\mathbf{x} \mid \boldsymbol{\theta}_g\right), \tag{1}$$

where $\pi_g \in (0,1]$, such that $\sum_{g=1}^{G} \pi_g = 1$, is the mixing proportion for component g, $f_x(\mathbf{x} \mid \boldsymbol{\theta}_g)$ is the density of a multivariate random variable $\mathbf{X}$ with parameters $\boldsymbol{\theta}_g$, and $\boldsymbol{\vartheta} = (\pi_1, \ldots, \pi_G, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G)$. Frequently, the finite Gaussian mixture model is used because of its mathematical tractability. This density is given by

$$f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g \phi\left(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\right), \tag{2}$$

where $\phi(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ is the density of a multivariate Gaussian random variable $\mathbf{X}$ with mean $\boldsymbol{\mu}_g$ and covariance matrix $\boldsymbol{\Sigma}_g$. An overview of model-based clustering is given by McNicholas [13].

Parsimonious Gaussian mixture models

The factor analysis model [14], assumes that a p-dimensional random vector $\mathbf{X}_i$ can be modelled using a q-dimensional vector of latent factors $\mathbf{U}_i$, where $q < p$. Factor analysis allows for a decrease in the number of parameters, which is useful in high-dimensional data cases. The model can be written as

$$\mathbf{X}_i = \boldsymbol{\mu} + \boldsymbol{\Lambda} \boldsymbol{U}_i + \boldsymbol{\epsilon}_i, \tag{3}$$

where $\boldsymbol{\Lambda}$ is a $p \times q$ matrix of factor loadings, the latent factors $\mathbf{U}_i \sim N(\mathbf{0}, \mathbf{I}_q)$ are independent, and the errors $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \boldsymbol{\Psi})$ are independently distributed and independent of the $\mathbf{U}_i$, where $\boldsymbol{\Psi}$ is a diagonal noise matrix with dimensions $p \times p$. Thus, $\mathbf{X}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi})$. In the mixture of factor analyzers (MFA) model, different factor analysis models are allowed in different regions of the feature subspace, using the density of a Gaussian mixture model with covariance structure $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g \boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}$ [15] or $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g \boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g$ [16]. The mixture of probabilistic principal components analysis (PPCA) model [17] is a special case of the MFA model from [16] in that it adds the assumption that the noise matrix is isotropic so that $\boldsymbol{\Psi}_g = \psi_g \mathbf{I}_p$. The parsimonious Gaussian mixture model (PGMM) family [18] allows combinations of the constraints: $\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$, $\boldsymbol{\Psi}_g = \boldsymbol{\Psi}$, and $\boldsymbol{\Psi}_g = \psi_g \mathbf{I}_p$ within the MFA model, resulting in a family of eight models.

Model-based biclustering

A recent review of biclustering on expression data by Pontes et al. [19] classifies the methods using various

taxonomies. One taxonomy is based on bicluster structure, specifically whether or not the genes and/or samples must be assigned to a bicluster (exhaustivity) and whether or not the genes and/or samples can be assigned to multiple biclusters (exclusivity). When considering blood biomarker discovery, an implicit property of the biomarker is that its expression profile is highly correlated between the blood and tissue of interest, and distinct from the rest of the tissues; indicating a unique biomarker for that tissue. Thus, the researcher would be interested in samples that are assigned to one bicluster only, in other words, non-overlapping column-exclusive biclusters. Examples of existing biclustering methods that adopt this property are plaid models developed by Lazzeroni and Owen [20], biclustering via Gibbs sampling developed by Sheng et al. [21], and Bayesian biclustering developed by Gu and Liu [22]. These are also examples of non metric-based probabilistic biclustering methods, based on another taxonomy provided in the review. The reader is referred to the review paper by Pontes et al. [19] for a structured and detailed discussion on the available biclustering methods.

Under the probabilistic framework, Martella et al. [23] propose a modified MFA technique for high-dimensional data for simultaneously clustering observations and variables. Variable cluster membership is represented by a binary row-stochastic matrix, which can be estimated via

$$\hat{\mathbf{\Lambda}}_g = \{\lambda_{gjl}\} = \begin{cases} 1 \text{ if } Q(\cdot, \lambda_{gjl} = 1) = \max_h Q(\cdot, \lambda_{gjh} = 1), \\ 0 \text{ otherwise,} \end{cases}$$

where $j = 1, \ldots, p$, $h, l = 1, \ldots, q$, $g = 1, \ldots, G$, and Q is the expected complete-data log-likelihood. We have $\mathbf{X}_i = \boldsymbol{\mu}_g + \mathbf{\Lambda}_g \mathbf{U}_{ig} + \boldsymbol{\epsilon}_{ig}$ with probability π_g. In this case, $\mathbf{U}_{ig} \sim N(\mathbf{0}, \mathbf{I}_q)$ and $\mathbf{X}_i \sim N(\boldsymbol{\mu}_g, \mathbf{\Lambda}_g \mathbf{\Lambda}_g' + \mathbf{\Psi}_g)$. This particular form of factor loadings matrix results in a block-diagonal covariance matrix which is especially suitable in the biclustering framework because it models the grouped nature of the variables. Additionally, it results in the non-overlapping biclusters that are useful in blood biomarker discovery. Constraining or not constraining the covariance parameters across clusters leads to a family of four models. This family will be referred to as MFABC from this point forward. The remainder of this paper describes a one-way supervised biclustering technique and its application to simulated and real data.

Methods

Covariance structure

To accommodate biclustering we set the factor loadings matrix to be binary row-stochastic. To allow for supervision along the variable dimension, we provide the structure of this matrix to the algorithm. In our gene expression analysis case, the variables are the samples, thus we are setting a relationship between the samples in the data set and providing it to the algorithm during initialization and take it as constant. Constraints can be imposed or not on $\mathbf{\Lambda}_g$, $\mathbf{\Psi}_g$, and $\mathbf{\Psi}_g = \psi_g \mathbf{I}_p$ to create a family of eight one-way-supervised Gaussian mixture models for biclustering (Table 1), which will be referred to as OSGaBi (**o**ne-way **s**upervised **Ga**ussian **bi**clustering) hereafter.

Parameter estimation

This section provides the mathematical details required to compute the parameter estimates for the eight members of the OSGaBi family, with a focus on the CUU model because it is the most appropriate model for the application presented previously and later in the Application section. The expectation-maximization (EM) algorithm [24] is an iterative procedure for computing the maximum likelihood estimates (MLE) when data are incomplete or treated as such. The EM algorithm is based on the complete-data, which consist of both observed and missing data. The algorithm begins with the expectation step (E-step), where the expected value of the complete-data log-likelihood (Q) is computed conditional on the current parameter estimates. In the maximization step (M-step),

Table 1 Properties of the OSGaBi family

Model nomenclature			Covariance structure ($\mathbf{\Sigma}_g$)	Covariance parameters
$\mathbf{\Lambda}_g = \mathbf{\Lambda}$	$\mathbf{\Psi}_g = \mathbf{\Psi}$	$\mathbf{\Psi}_g = \psi_g \mathbf{I}_p$		
C	C	C	$\mathbf{\Lambda}\mathbf{\Lambda}' + \psi \mathbf{I}_p$	1
C	C	U	$\mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}$	p
C	U	C	$\mathbf{\Lambda}\mathbf{\Lambda}' + \psi_g \mathbf{I}_p$	G
C	U	U	$\mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}_g$	Gp
U	C	C	$\mathbf{\Lambda}_g \mathbf{\Lambda}_g' + \psi \mathbf{I}_p$	1
U	C	U	$\mathbf{\Lambda}_g \mathbf{\Lambda}_g' + \mathbf{\Psi}$	p
U	U	C	$\mathbf{\Lambda}_g \mathbf{\Lambda}_g' + \psi_g \mathbf{I}_p$	G
U	U	U	$\mathbf{\Lambda}_g \mathbf{\Lambda}_g' + \mathbf{\Psi}_g$	Gp

The nomenclature, covariance structure, and number of covariance parameters for each member of the OSGaBi family. C, constrained; U, unconstrained

Q is maximized with respect to the model parameters. These two steps are repeated until convergence.

The alternating expectation-conditional maximization (AECM) algorithm [25] incorporates a series of conditional maximization (CM) steps instead of a single M-step and also allows for different specification of the complete-data at each stage. This algorithm is used for parameter estimation for the MFA model, the PGMM family, and the MFABC family. It will also be used for the OSGaBi family.

For convenience, the following notation is adopted. We denote the observed data as $\mathbf{x}$ and the unobserved latent parameters as $\mathbf{U}_i = (\mathbf{U}_{i1}, \ldots, \mathbf{U}_{iG})$. We denote the missing group memberships as $\mathbf{z}_i$, where

$$z_{ig} = \begin{cases} 1 & \text{if observation } i \text{ belongs to component } g, \\ 0 & \text{otherwise,} \end{cases}$$

for $i = 1, \ldots, n, g = 1, \ldots, G$.

In the first cycle of the AECM algorithm, $(\mathbf{x}_i, \mathbf{z}_i)$ are the complete-data, where $i = 1, \ldots, n$. During the CM-step, π_g and $\boldsymbol{\mu}_g$ are updated. During the E-step, the z_{ig} are replaced by their expected values

$$\mathbb{E}[Z_{ig} \mid \hat{\pi}_g, \hat{\boldsymbol{\mu}}_g, \hat{\boldsymbol{\Lambda}}_g, \hat{\boldsymbol{\Psi}}_g] = \frac{\hat{\pi}_g \phi(\mathbf{x}_i \mid \hat{\boldsymbol{\mu}}_g, \hat{\boldsymbol{\Lambda}}_g, \hat{\boldsymbol{\Psi}}_g)}{\sum_{h=1}^{G} \hat{\pi}_h \phi(\mathbf{x}_i \mid \hat{\boldsymbol{\mu}}_h, \hat{\boldsymbol{\Lambda}}_h, \hat{\Psi}_h)} =: \hat{z}_{ig}$$

leading to the calculation of the expected value of the complete-data log-likelihood, Q_1. In the CM-step, Q_1 is maximized to give

$$\hat{\pi}_g = \frac{n_g}{n}$$

and

$$\hat{\boldsymbol{\mu}}_g = \frac{\sum_{i=1}^{n} \hat{z}_{ig} \mathbf{x}_i}{n_g},$$

where $n_g = \sum_{i=1}^{n} \hat{z}_{ig}$.

In the second cycle of the AECM algorithm, $(\mathbf{x}_i, \mathbf{z}_i, \mathbf{U}_i)$ are the complete-data. During this CM-step, $\boldsymbol{\Psi}_g$ is updated. During this E-step, z_{ig} are replaced by $\hat{z}_{ig}$ and $\mathbf{U}_{ig}$ and $\mathbf{U}_{ig}\mathbf{U}_{ig}'$ are replaced by

$$\begin{aligned} \mathbb{E}\left[\mathbf{U}_{ig} \mid \mathbf{x}_i, \boldsymbol{\mu}_g, \boldsymbol{\Lambda}, \boldsymbol{\Psi}_g\right] &= \boldsymbol{\beta}_g \sum_{i=1}^{n} \hat{z}_{ig}(\mathbf{x}_i - \boldsymbol{\mu}_g), \\ \mathbb{E}\left[\mathbf{U}_{ig}\mathbf{U}_{ig}' \mid \mathbf{x}_i, \boldsymbol{\mu}_g, \boldsymbol{\Lambda}, \boldsymbol{\Psi}_g\right] &= \mathbf{I}_q - \boldsymbol{\beta}_g \boldsymbol{\Lambda} \\ &\quad + \boldsymbol{\beta}_g \sum_{i=1}^{n} \hat{z}_{ig}(\mathbf{x}_i - \boldsymbol{\mu}_g)(\mathbf{x}_i - \boldsymbol{\mu}_g)' \boldsymbol{\beta}_g' \end{aligned} \tag{4}$$

respectively, where $\boldsymbol{\beta}_g = \boldsymbol{\Lambda}'(\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}_g)^{-1}$ for model CUU, to allow for the calculation of Q_2. In the CM-step, the maximization of Q_2 is specific for each model. Considering the CUU model,

$$\begin{aligned} Q_2(\boldsymbol{\Lambda}, \boldsymbol{\Psi}_g) = C + \sum_{g=1}^{G} \frac{n_g}{2} \Big[&\log|\boldsymbol{\Psi}_g^{-1}| - \operatorname{tr}\left\{\boldsymbol{\Psi}_g^{-1}\mathbf{S}_g\right\} \\ &+ 2\operatorname{tr}\left\{\boldsymbol{\Psi}_g^{-1}\boldsymbol{\Lambda}\boldsymbol{\beta}_g\mathbf{S}_g\right\} \\ &- \operatorname{tr}\left\{\boldsymbol{\Psi}_g^{-1}\boldsymbol{\Lambda}\boldsymbol{\Theta}_g\boldsymbol{\Lambda}'\right\}\Big], \end{aligned} \tag{5}$$

where C is a constant with respect to the unknown parameters, $\mathbf{S}_g = \frac{1}{n_g}\sum_{i=1}^{n} \hat{z}_{ig}(\mathbf{x}_i - \boldsymbol{\mu}_g)(\mathbf{x}_i - \boldsymbol{\mu}_g)'$, and $\boldsymbol{\Theta}_g = \mathbf{I}_q - \boldsymbol{\beta}_g\boldsymbol{\Lambda} + \boldsymbol{\beta}_g\mathbf{S}_g\boldsymbol{\beta}_g'$.

The following score function is obtained when differentiating Q_2 with respect to $\boldsymbol{\Psi}_g$:

$$\begin{aligned} S(\boldsymbol{\Lambda}, \boldsymbol{\Psi}_g) &= \frac{\delta Q}{\delta \boldsymbol{\Psi}_g^{-1}} \\ &= \sum_{g=1}^{G} \frac{n_g}{2}\Big[\boldsymbol{\Psi}_g - \mathbf{S}_g + 2\boldsymbol{\Lambda}\boldsymbol{\beta}_g\mathbf{S}_g - \boldsymbol{\Lambda}\boldsymbol{\Theta}_g\boldsymbol{\Lambda}'\Big]. \end{aligned} \tag{6}$$

Now, setting $S(\boldsymbol{\Lambda}, \hat{\boldsymbol{\Psi}}_g) = 0$ and solving gives the estimate

$$\hat{\boldsymbol{\Psi}}_g = \operatorname{diag}\left\{\mathbf{S}_g - 2\boldsymbol{\Lambda}\boldsymbol{\beta}_g\mathbf{S}_g + \boldsymbol{\Lambda}\boldsymbol{\Theta}_g\boldsymbol{\Lambda}'\right\}.$$

The parameter estimates for the remaining seven models are derived similarly and are provided in the Additional file 1 titled `OSGaBi_MWong_appendix.pdf`.

When running the AECM algorithm, utilizing the Woodbury identity [26] avoids inverting any non-diagonal $p \times p$ matrices that may be singular for $p \gg n$. Suppose an $n \times n$ matrix $\mathbf{A}$, an $n \times q$ matrix $\mathbf{H}$, a $q \times q$ matrix $\mathbf{C}$, and a $q \times n$ matrix $\mathbf{V}$. The Woodbury identity states that

$$(\mathbf{A} + \mathbf{HCV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{H}(\mathbf{C}^{-1} + \mathbf{V}\mathbf{A}^{-1}\mathbf{H})^{-1}\mathbf{V}\mathbf{A}^{-1}. \tag{7}$$

Specifically for the AECM algorithm, setting $\mathbf{H} = \boldsymbol{\Lambda}$, $\mathbf{V} = \boldsymbol{\Lambda}'$, $\mathbf{A} = \boldsymbol{\Psi}$, and $\mathbf{C} = \mathbf{I}_q$ results in

$$(\boldsymbol{\Psi} + \boldsymbol{\Lambda}\boldsymbol{\Lambda}')^{-1} = \boldsymbol{\Psi}^{-1} - \boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}(\mathbf{I}_q + \boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}. \tag{8}$$

Now, instead of inverting the $p \times p$ covariance matrix on the left side of Eq. 8, only the diagonal and $q \times q$ matrices on the right side need to be inverted. With gene expression data where $q \ll p$, this identity provides a major computational advantage. Another useful identity is for calculating

the determinant of the covariance matrix in the AECM algorithm:

$$|\boldsymbol{\Psi} + \boldsymbol{\Lambda}\boldsymbol{\Lambda}'| = \frac{|\boldsymbol{\Psi}|}{|\mathbf{I}_q - \boldsymbol{\Lambda}'(\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi})^{-1}\boldsymbol{\Lambda}|}.$$

Component membership

The predicted biclustering for each member of the OSGaBi family is given by the maximum *a posteriori* (MAP) classification for the observations and the classifications originally provided for the variables. That is, the posterior predicted component membership of observation (i.e., gene) i is the value of g for which $\hat{z}_{ig}$ is greatest. In the biological sense, this will identify which gene belongs to which subset, implying that the genes in each subset are related in some way. Component membership of variable (i.e., sample) j is already provided as $\boldsymbol{\Lambda}_g$ at the beginning of the algorithm, specifically

$$\boldsymbol{\Lambda}_g = \{\lambda_{gjl}\} = \begin{cases} 1 & \text{if variable } j \text{ belongs to cluster } l, \\ 0 & \text{otherwise,} \end{cases}$$

for $j = 1, \ldots, p$, $l = 1, \ldots, q$, and $g = 1, \ldots, G$. In the biological view, we know a priori that a certain set of samples are/should be related to each other, which is uncorrelated to another set of samples. A concrete example of how component membership is applied in microarray gene expression analysis is presented in the Application section.

Convergence and model selection

Convergence of the AECM algorithm is determined using the Aitken's acceleration [27] to estimate the asymptotic maximum of the log-likelihood at each iteration of the AECM algorithm for a specific number of components and a specific number of factors, as described in [28]. The Aitken's acceleration at iteration t is

$$a^{(t)} = \frac{l^{(t+1)} - l^{(t)}}{l^{(t)} - l^{(t-1)}},$$

where l corresponds to the respective log-likelihood. The asymptotic estimate of the log-likelihood at iteration $t+1$ is

$$l_\infty^{(t+1)} = l^{(t)} + \frac{1}{1 - a^{(t)}}\left(l^{(t+1)} - l^{(t)}\right)$$

[29]. The stopping criterion $l_\infty^{(t+1)} - l^{(t)} < \epsilon$ [30], where $\epsilon = 0.1$, is used and provided that the difference is positive [13]. The Bayesian information criterion (BIC) [31] is used to choose the best member of the proposed OSGaBi family with respect to the model and number of components, G.

Results

Simulation studies

Simulation studies were carried out to validate the proposed biclustering algorithm. The adjusted Rand index (ARI) [32] was used to evaluate the performance of the algorithm in recovering biclusters from the simulated data. Specifically, $\mathbf{z}_i$ was compared with $\hat{\mathbf{z}}_i$ after convergence was attained. Model selection was done via the BIC as previously described, although it can be noted that the integrated completed likelihood (ICL) [33] and Akaike information criterion (AIC) [34] were used as comparison and produced the same outcomes. The parameters and resulting data sets for the following simulation studies are found in the (Additional file 1: Supplementary files).

Simulated data were generated with $G = 2$, 3, and 4 clusters for observations and $q = 2$ clusters for variables. This resulted in 4, 6, and 8 biclusters, respectively. Four cases were examined: low, medium, and high variance coupled with good cluster separation, and high variance coupled with relatively close clusters. For each case, 100 data sets were generated, where each set had $p = 8$ variables and $n = 200, 300, 400$ observations (for $G = 2, 3, 4$, respectively) and was randomly generated from the same Gaussian distribution. Examples of heatmaps for each of the CUC cases visibly indicate that there are distinct biclusters in the simulated data (Fig. 1). To reflect the one-way supervised nature of the algorithm, the true $\boldsymbol{\Lambda}$ was provided. Twenty random starts were used for each run of the algorithm. Table 2 presents the results from these four simulation studies for the CUU and CUC models. It lists the average number of components selected, the most frequently chosen model, and the average ARI when fitting $G = 2, \ldots, 10$ observation clusters. Because the algorithm was sometimes overfitting for the number of components based on the model it chose, another analysis was included to show the average ARI when the number of clusters was known (i.e., $G = 2$, 3, 4, depending on the case). These results are shown in the last column. The CUU and CUC models are focused on because they are the most probable cases in real-life scenarios, and additionally, they are the models most frequently selected when the number of clusters was known (results not shown).

For completeness, simulation studies were conducted on the remaining six OSGaBi models using simulated data with medium variance and good cluster separation and with the same properties as that used for the CUU and CUC models. As the true $\boldsymbol{\Lambda}_g$ or $\boldsymbol{\Lambda}$ was provided, it implied that for models with unconstrained $\boldsymbol{\Lambda}$ (i.e., models UUU, UUC, UCU, and UCC), G was known because the $\boldsymbol{\Lambda}$ for each component would have been provided. Table 3 presents the average ARI, most frequently chosen model, and average number of clusters selected from this simulation study for each of the remaining six models when fitting $G = 2, \ldots, 10$ clusters for observations.

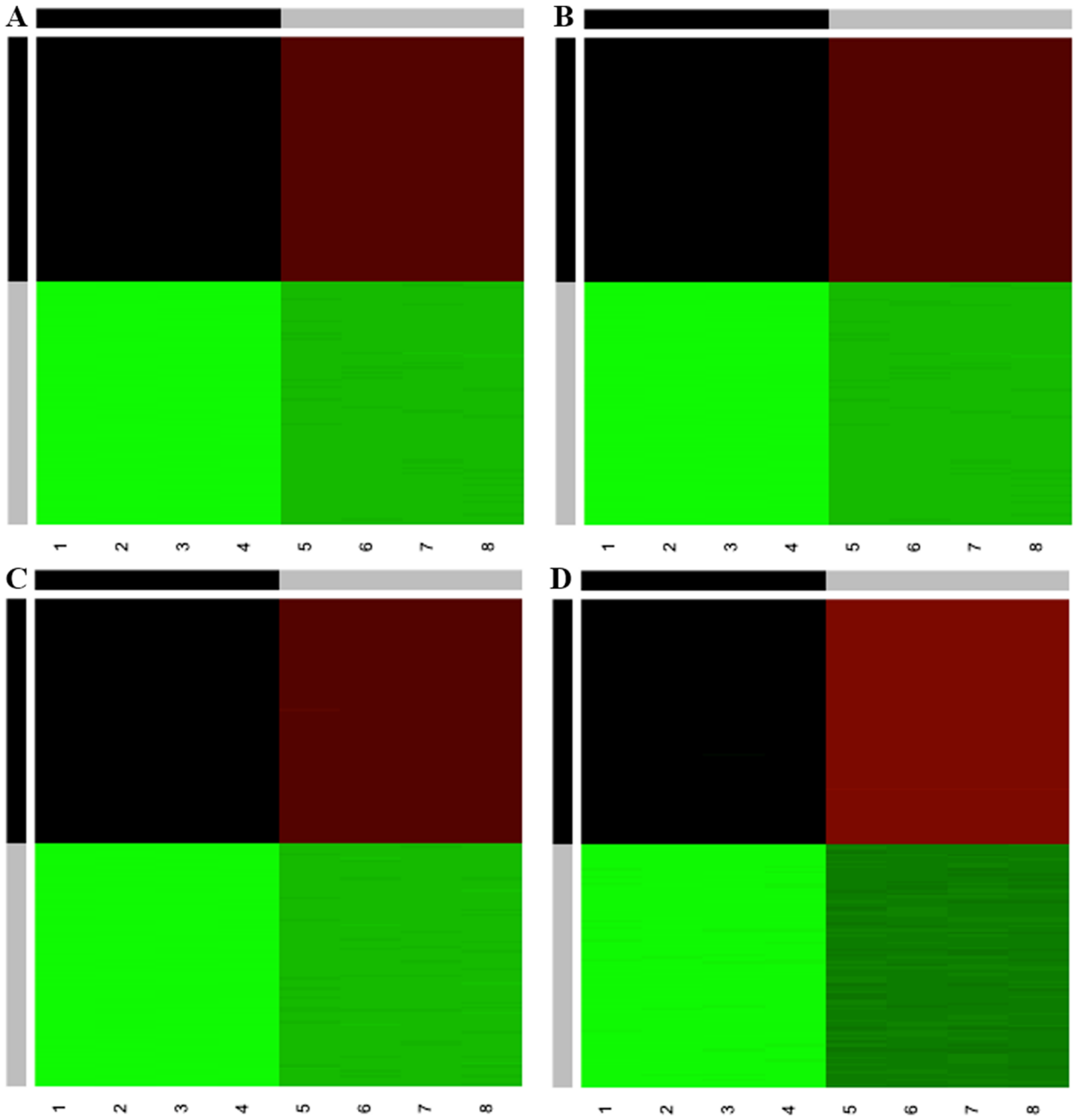

Fig. 1 Heatmaps of simulated data for CUC model. Examples of heatmaps of the four types of simulated data used for the CUC model: low variance and good cluster separation (**a**), medium variance and good cluster separation (**b**), high variance and good cluster separation (**c**), and high variance with relatively close clusters (**d**)

Because the algorithm was once again sometimes overfitting for the number of components based on the model selected, this final analysis was included to show the algorithm's performance when the number of clusters was fixed to $G = g_{\text{known}}$, where g_{known} represents the number of observation clusters the data was generated from. The last column of the table presents the corresponding average ARI when fixing G for the CCU and CCC models.

It is important to note that although the algorithm was overfitting for the number of components based on the model selected, the majority of the time the original components were simply being broken into smaller components. A classification table from one of the simulation results illustrates the very common occurrence (Table 4). In this specific result, Cluster 1 was broken up into three components by the algorithm, resulting in a

Table 2 Simulation study results for model CUU

	$G = 2, \ldots, 10$			$G = g_{\text{known}}$
Case	Average G	Most chosen model	Average ARI	Average ARI
CUC, $g_{\text{known}} = 2$				
Low var, good cluster sep	2.5 (0.8)	CUU	0.955 (0.083)	1.0 (0.0)
Mid var, good cluster sep	2.4 (0.8)	CUU	0.955 (0.094)	1.0 (0.0)
High var, good cluster sep	4.0 (1.0)	CUC	0.708 (0.106)	1.0 (0.0)
High var, close clusters	6.4 (1.2)	CUC	0.502 (0.135)	1.0 (0.0)
CUU, $g_{\text{known}} = 2$				
Low var, good cluster sep	2.4 (0.7)	CUU	0.969 (0.072)	1.0 (0.0)
Mid var, good cluster sep	2.5 (0.8)	CUU	0.964 (0.071)	1.0 (0.0)
High var, good cluster sep	4.0 (1.0)	CUC	0.705 (0.103)	1.0 (0.0)
High var, close clusters	6.4 (1.3)	CUC	0.485 (0.138)	1.0 (0.0)
CUC, $g_{\text{known}} = 3$				
Low var, good cluster sep	3.5 (0.7)	CUU	0.981 (0.039)	1.0 (0.0)
Mid var, good cluster sep	3.4 (0.7)	CUU	0.984 (0.033)	1.0 (0.0)
High var, good cluster sep	5.1 (1.0)	CUC	0.864 (0.081)	1.0 (0.0)
High var, close clusters	8.8 (1.1)	CCC	0.601 (0.066)	1.0 (0.0)
CUU, $g_{\text{known}} = 3$				
Low var, good cluster sep	3.5 (0.6)	CUU	0.984 (0.028)	1.0 (0.0)
Mid var, good cluster sep	3.4 (0.7)	CUU	0.975 (0.050)	1.0 (0.0)
High var, good cluster sep	5.0 (1.1)	CUC	0.866 (0.079)	1.0 (0.0)
High var, close clusters	8.8 (1.0)	CUC	0.590 (0.070)	1.0 (0.0)
CUC, $g_{\text{known}} = 4$				
Low var, good cluster sep	4.4 (0.7)	CUU	0.989 (0.254)	1.0 (0.0)
Mid var, good cluster sep	4.3 (0.5)	CUU	0.992 (0.020)	1.0 (0.0)
High var, good cluster sep	6.2 (1.0)	CUC	0.887 (0.048)	1.0 (0.0)
High var, close clusters	9.7 (0.5)	CUC	0.658 (0.045)	1.0 (0.0)
CUU, $g_{\text{known}} = 4$				
Low var, good cluster sep	4.4 (0.9)	CUU	0.989 (0.031)	1.0 (0.0)
Mid var, good cluster sep	4.4 (0.7)	CUU	0.989 (0.024)	1.0 (0.0)
High var, good cluster sep	4.6 (0.8)	CUU	0.970 (0.048)	1.0 (0.0)
High var, close clusters	9.8 (0.5)	CUC	0.653 (0.046)	1.0 (0.0)

Average ARI, most frequently chosen model, and the number of observation clusters selected for the CUU and CUC models using simulated data with low, medium, and high variance (var) with good cluster separation (sep), and high variance with relatively close clusters when fitting $G = 2, \ldots, 10$ observation clusters using 100 data sets and 20 random starts. The last column presents the ARI when fixing $G = g_{\text{known}}$, where g_{known} represents the number of observation clusters the data was generated from. Values in brackets represent the respective standard deviation

total of four components. The final column of Tables 2 and 3 provide further evidence because once the algorithm is provided the correct number of components, the ARI become perfect or near perfect.

Application

Rat data

We present the biclustering results from Affymetrix oligonucleotide array data from a nutritional and pharmaceutical intervention in diabetic rats . This study consisted of five male lean control rats and five male Zucker diabetic fatty (ZDF) rats, which are genetically predisposed to developing diabetes. Details regarding the original rat study are described in Beaudoin et al. [35]. From each animal, tissue was extracted from various tissue depots, including the liver and red tibialis anterior (red TA, a type of muscle). Blood was also extracted, resulting in a total of 30 samples. RNA was extracted from these samples and used for the subsequent microarray gene expression analysis. Pre-processed data can be found on Gene Expression Omnibus (GEO) [36], accession number GSE93402 (blood), GSE93403 (liver), and GSE93406

Table 3 Simulation study results for the other six models

	$G = 2, \ldots, 10$			$G = g_{\text{known}}$
Model	Average G	Most chosen model	Average ARI	Average ARI
$g_{\text{known}} = 2$				
UUU	2	UUC	1.0 (0.0)	NA
UUC	2	UUC	1.0 (0.003)	NA
UCU	2	UCC	0.960 (0.028)	NA
UCC	2	UCC	0.961 (0.025)	NA
CCU	2.3 (0.7)	CCU	0.971 (0.098)	1.0 (0.0)
CCC	2.4 (1.2)	CCU	0.959 (0.113)	1.0 (0.0)
$g_{\text{known}} = 3$				
UUU	3	UUC	1.0 (0.0)	NA
UUC	3	UUC	1.0 (0.0)	NA
UCU	3	UCC	1.0 (0.0)	NA
UCC	3	UCC	1.0 (0.0)	NA
CCU	4.0 (1.3)	CCU	0.936 (0.095)	1.0 (0.0)
CCC	4.3 (1.4)	CCU	0.915 (0.107)	1.0 (0.0)
$g_{\text{known}} = 4$				
UUU	4	UUC	1.0 (0.0)	NA
UUC	4	UUC	1.0 (0.0)	NA
UCU	4	UCC	1.0 (0.0)	NA
UCC	4	UCC	1.0 (0.0)	NA
CCU	5.1 (1.3)	CCU	0.958 (0.057)	1.0 (0.055)
CCC	5.1 (1.1)	CCU	0.960 (0.052)	1.0 (0.029)

Average ARI, most chosen model, and the average number of observation clusters selected for the remaining six OSGaBI models using simulated data with medium variance and good cluster separation when fitting $G = 2, \ldots, 10$ observation clusters using 100 data sets and 20 random starts. The last column presents the ARI when fixing $G = g_{\text{known}}$, where g_{known} represents the number of observation clusters the data was generated from, for the CCU and CCC models. Values in brackets represent the respective standard deviation

(red TA). After pre-processing using the `affy` and `oligo` packages [37, 38] respectively for R Bioconductor [39, 40] respectively, $n = 8801$ genes remained. We worked with the top 2000 differentially expressed genes between the red TA and liver ($p < 0.01$). For this analysis, we set the genes as the observations ($n = 2000$) and the samples as the variables ($p = 30$).

The goal of the biclustering analysis was to identify sets of genes within the blood that possess similar expression profiles within the distinct tissues. Thus, we aimed to match biclusters containing genes that had similar expression profiles that were unique for blood and a specific tissue type. We focus here on genes with similar expression profiles between blood and liver. Downstream, these candidate genes can be tested to determine if they can function as blood biomarkers of metabolic status in individuals in different contexts (i.e., response to interventions, different disease states, etc.); however, this subsequent analysis goes beyond the scope of the present article.

Table 4 An example of a classification table from one of the simulation results

	True			
	1	2	3	4
Estimated 1	56	32	12	0
2	0	0	0	100

Although the algorithm was overfitting for the number of components based on the model selected, the majority of the time the original components were simply being broken into smaller components

We constrained the structure of $\mathbf{\Lambda}_g$ because we knew the relationships required among the three sample types. Specifically, we wanted correlated expression between blood and liver only, implying that expression between blood and red TA were uncorrelated and expression between liver and red TA were uncorrelated as well. The other (extraneous) relationship characterized by the block-diagonal covariance matrix was the correlated nature of the expression strictly among the liver samples. Consequently, $q = 2$ for the number of variable clusters (i.e., the two relationships described previously). Sample types were constant across all components, i.e.,

$\mathbf{\Lambda}_g = \mathbf{\Lambda}$, and thus we limited the algorithm to fit model CUU and CUC. These two Λ-constrained models were chosen based on the results from the simulation studies previously mentioned. We normalized the data and fitted the range of $G = 2, \ldots, 30$.

The BIC selected a CUU model with $G = 19$ observation clusters for the blood-liver analysis. As seen from the heatmaps before and after biclustering and subsequent rearranging, there were definitive biclusters in the data (Fig. 2). We inputted the gene lists for each of the 19 biclusters into the online functional annotation tool DAVID (Database for Annotation, Visualization and Integrated Discovery) [41, 42] to elucidate potential biological processes that were dominant in each bicluster. DAVID functional annotation results indicated that the largest proportions of genes in the blood-liver biclusters had roles in protein metabolic and modification processes, carboxylic metabolic process, oxaloacid metabolic process, and intracellular signal transduction (biological processes as defined by the Gene Ontology Consortium, [43]), all biological processes of which have previously been shown to have an involvement in diabetes and obesity, and some processes within the liver [44–50]. These processes accounted for approximately 20–43% of the genes in the various biclusters and were all statistically significantly enriched ($p < 0.05$). There is also a general inference that insulin resistance occurs at different times in insulin sensitive tissues such as muscle and liver [51, 52]; therefore, it is not surprising that the expression profiles between the liver and red TA were not similar. Additionally, it has been previously established that the peripheral blood transcriptome reflects changes in various tissues throughout the body [3], a property that is illustrated in the biclusters of interest.

Human data

The second data set we analyzed is another Affymetrix oligonucleotide array retrieved from GEO, accession number GSE1133. The original study aimed to profile 79 human and 61 mouse tissues in terms of their transcriptomes under normal conditions [53]. Here, we focus on the human arrays, specifically the tissues related to the immune system (20 tissue types) and the brain (16 tissue types), and also whole blood, for a total of 37 tissue types. Each tissue had two replicates, giving a total of 74 samples. After pre-processing using the same methods described for the rat data and removing genes without Entrez gene identifiers, $n = 3867$ genes remained, of which 2148 genes were differentially expressed between brain and immune tissues ($p < 0.01$). Similar to the rat data, we set the genes

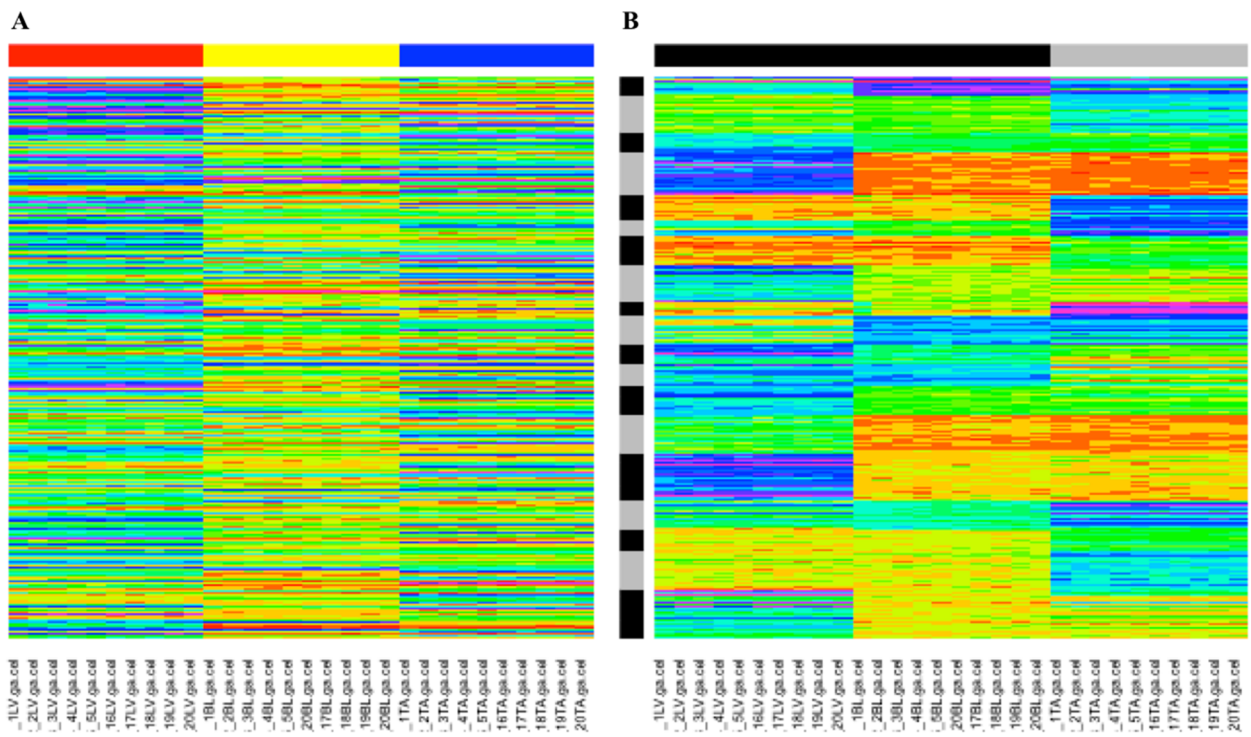

Fig. 2 Heatmaps of the rat data. Heatmap of the rat data before biclustering (**a**). The *red*, *yellow*, and *blue bars* along the columns and represent liver, blood, and red TA samples, respectively. Heatmap of the rat data after biclustering and subsequently rearranging the rows so that the observation clusters were contiguous (**b**). *Black* and *grey bars* along the columns and rows simply represent the presence of the clusters and do not indicate relationships between them. $G = 19$ for the observation clusters and along the columns, $q_1 = q_2 = 2$

as the observations ($n = 2148$) and the samples as the variables ($p = 74$).

The goal of this biclustering analysis was to identify sets of genes within the blood that possess similar expression within the distinct groups of tissues. Thus, we aimed to match biclusters containing genes that had similar expression that were unique for blood and a specific group of tissues. We focus here on genes with similar expression between blood and immune tissues. Subsequent work can involve determining which of these candidate genes can function as blood biomarkers of normal immune function in individuals.

Similar to the rat data, we constrained the structure of $\mathbf{\Lambda}_g$ because we knew the relationships required among the three sample types. Specifically, we wanted correlated expression only between blood and immune tissues. This implied that expression between blood and brain tissues were uncorrelated, and expression between immune and brain tissues were uncorrelated as well. The other (extraneous) relationship characterized by the block-diagonal covariance matrix was the correlated nature of the expression strictly among the immune tissue samples. Consequently, $q = 2$ for the number of variable clusters (i.e., the two relationships described previously). Samples were constant across all components, i.e., $\mathbf{\Lambda}_g = \mathbf{\Lambda}$, and thus we again limited the algorithm to fit model CUU and CUC. We normalized the data and fitted models in the range of $G = 2, \ldots, 30$.

The BIC selected a CUU model with $G = 10$ observation clusters for the blood-immune analysis. As seen from the heatmaps before and after biclustering and subsequent rearranging, there were again definitive biclusters in the data (Fig. 3). DAVID functional annotation results indicated that the largest portion of genes in each bicluster had roles in the nucleobase-containing small molecule metabolic process, macromolecule metabolic process, microtubule-based process, microtubule cytoskeletal organization, response to DNA damage stimulus, and transmembrane transport; all biological processes that have been linked to immune responses [54–57]. These processes accounted for anywhere between 4 to 51% of the genes in the various biclusters, and were all statistically significantly enriched ($p < 0.05$). Furthermore, blood acts as a transporter for the immune system by transporting immune cells throughout the body, thus blood can provide an extensive view of the immune status of an individual [58]. This property is reflected in the biclusters of interest because there is a

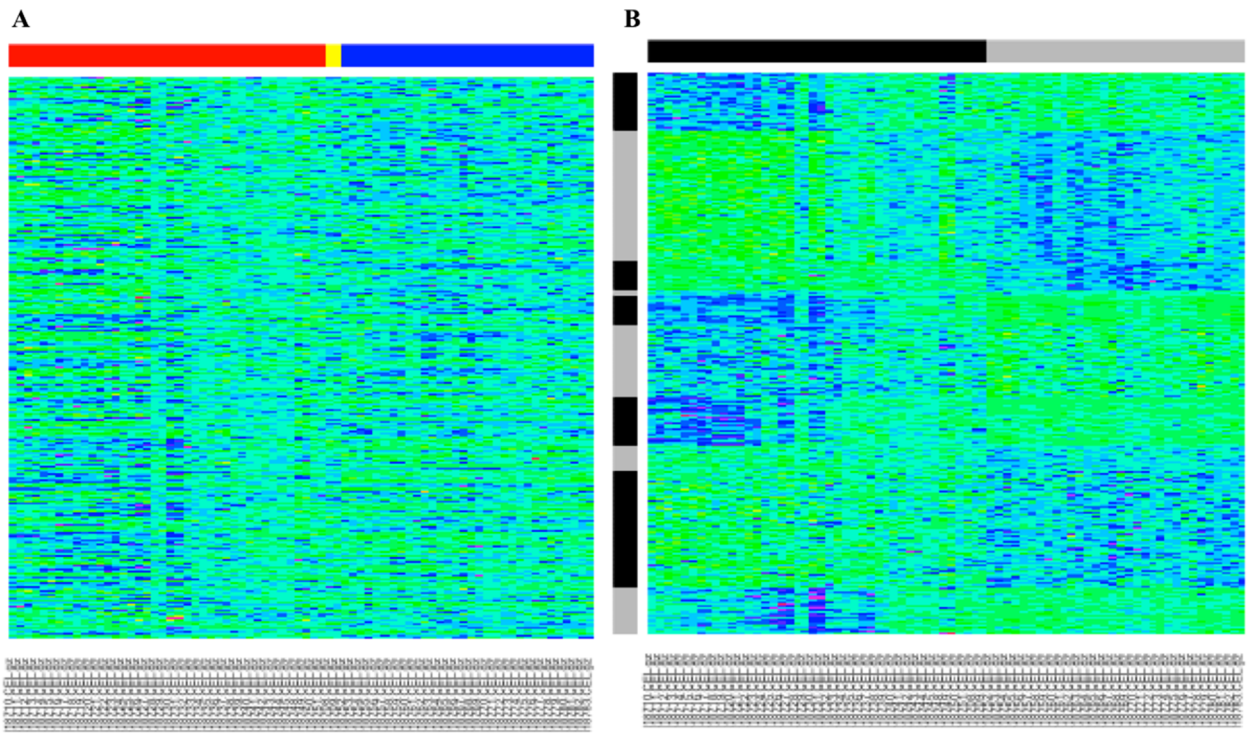

Fig. 3 Heatmaps of the human data. Heatmap of the human data before biclustering (**a**). The *red*, *yellow*, and *blue bars* along the columns and represent immune tissues, whole blood, and brain tissues respectively. Heatmap of the human data after biclustering and subsequently rearranging the rows so that the observation clusters were contiguous (**b**). *Black* and *grey bars* along the columns and rows simply represent the presence of the clusters and do not indicate relationships between them. $G = 10$ for the observation clusters and along the columns, $q_1 = q_2 = 2$

correlation among the expression between the blood and the immune tissues.

Discussion

One specific taxonomy of biclustering methods for gene expression data aims to retrieve non-overlapping biclusters characterized by one specific sample type (in this case, "sample" could refer to a type of treatment, tissue, disease state, etc.) along the variable dimension, as reviewed in Pontes et al. [19]. This is useful in applications such as disease subtype discovery, where the focus is to elucidate the various disease subtypes based on the genes. Conversely, in blood biomarker discovery, knowledge of the types of samples a priori is required and the focus is on the relationships between those sample types based on the genes, which is where one-way supervised biclustering is able to play a role. Two inherent criteria of blood biomarkers are that there is 1) a correlation between blood and the tissue of interest and 2) no correlation between those two sample types and other tissues. The second criteria is enforced by including other tissues into the biclustering analysis so that the condition can be set in conjunction with the first criteria. The relationships required are satisfied through the use of one-way supervision to explicitly determine the relationship between blood and the various tissues. To the best of our knowledge, biclustering methods currently available under the taxonomy of non-overlapping biclusters do not provide the option of one-way supervision along the variable dimension to aid in applications such as blood biomarker discovery.

Another advantage of approaching tissue-specific blood biomarker discovery through the use of biclustering is the ability to identify groups of genes that are potentially related to each other through their biological pathways. Commonly, correlation analysis between blood and a tissue is conducted using the available gene list in its entirety, e.g. [59], consequently not revealing any information about genes related by biological pathways that a cluster analysis would provide. In our OSGaBi family, setting the variable clusters labels and subsequently biclustering conditional on this information allows us to handle this limitation of correlation analysis.

Simulation study results show that models with values of G that are too high are sometimes selected, and this problem becomes more pronounced for high variance. While the BIC has been shown to be unreliable in higher dimensions, e.g. [60] — and this may suggest that further research on an optimal model selection criteria for this family of biclustering models is warranted — it is quite possible that the selection of larger values of G is simply a result of lack of concentration around the modes at higher variances. The inclusion of results for fixed G follows [61] and [62], who carried out mixture model analysis of gene expression data by treating G as fixed and known. Note that, in [23] where the binary row-stochastic factor loadings matrix is a property of their MFABC family, the authors report simulation results but do not mention the model selection criterion or the range of number of observation clusters fitted; therefore, it is not known if the authors treated G as fixed. Conversely, the authors mention the use of the BIC and AIC for model selection in their real data study with gene expression data, supporting the use of the BIC for our analyses until the optimal model selection criteria is determined.

Future work will also aim to compare performance of the OSGaBi family to that of other model-based biclustering algorithms capable of detecting non-overlapping clusters and allowing for one-way supervision. Current methods are available for the former (as mentioned previously), but do not allow for the latter criteria. This limitation in the existing methods makes it difficult to compare the genes that are found in the biclusters to those found using the OSGaBi family since they do not always correspond to the intended subset of variables.

We have presented biclustering results using the OSGaBi family on two real microarray gene expression data sets. The first one was a previously unpublished rat microarray gene expression data set, where identified biclusters corresponded to genes whose expression profiles were correlated between liver and blood (and not between red TA and blood, or liver and red TA). Identified biclusters were enriched in genes related to biological processes known to play a role in insulin resistance and obesity in a tissue-specific manner. The second data set was a subset of a microarray gene expression data set from the GEO database that aimed to profile the human transcriptome under normal conditions. In this analysis, identified biclusters corresponded to genes whose expression correlated between immune tissues and blood (and not between brain tissues and blood, or immune and brain tissues). Identified biclusters contained genes related to biological processes previously associated with the immune system. Although further biological experimental analysis and interpretation need to be conducted to determine the best candidate gene(s) in both preliminary analyses, the initial results show promise in using the OSGaBi biclustering family for discovering novel blood biomarkers to act as surrogate tissue material in the maintenance of health and the prevention of disease.

Conclusions

A family of parsimonious Gaussian mixture models for the biclustering of gene expression data has been proposed. These models work in a one-way-supervised fashion in that the variable labels are known. The binary and row-stochastic factor loadings matrix results in a block-diagonal covariance matrix, which can be a useful property in biclustering applications for dictating the

relationships between the variables. A promising application for our method is in the discovery of novel peripheral blood biomarkers for use as surrogate biopsy material.

Additional file

Additional file 1: Supplementary files. (PDF 90 kb)

Abbreviations

AECM: Alternating expectation-conditional maximization; AIC: Akaike information criterion; ARI: Adjusted rand index; BIC: Bayesian information criterion; DAVID: Database for annotation: visualization and integrated discovery; EM: Expectation-maximization; GEO: Gene expression omnibus; ICL: Integrated completed likelihood; MAP: Maximum *a posteriori*; MFA: Mixtures of factor analyzers; MFABC: Mixtures of factor analyzers for biclustering; MLE: Maximum likelihood estimates; OSGaBi: One-way supervised Gaussian biclustering; PGMM: Parsimonious Gaussian mixture model; PPCA: Probabilistic principal components analysis; TA: Tibialis anterior; ZDF: Zucker diabetic fatty

Acknowledgements

Not applicable.

Funding

This research was funded by an Ontario Early Researcher Award (PDM) and a CIHR Catalyst Grant in Environment, Genes and Chronic Disease (DMM, PDM). This work was partially supported by the Canada Research Chairs program (PDM).

Availability of data and material

Pre-processed rat microarray data can be found on GEO:
`GSE93402` blood
`GSE93403` liver
`GSE93406` red TA

Authors' contributions

PDM initiated and supervised the project. MHTW developed, implemented, and tested the algorithm and drafted the manuscript. MHTW conducted the original rat microarray study mentioned in the Application section. DMM provided biological insight and supervised the original rat microarray study. All three authors wrote and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

For the rat data set in the Application section, all protocols followed Canadian Council on Animal Care guidelines and were approved by the Animal Care Committee at the University of Guelph.

Author details

[1]Department of Mathematics and Statistics, McMaster University, L8S 4L8 Hamilton, ON, Canada. [2]Department of Human Health and Nutritional Sciences, University of Guelph, N1G 2W1 Guelph, ON, Canada.

Received: 5 August 2016 Accepted: 24 February 2017
Published online: 04 March 2017

References

1. Offit K. Personalized medicine: new genomics, old lessons. Hum Genet. 2011;130(1):3–14.
2. Ghosh D, Poisson LM. "Omics" data and levels of evidence for biomarker discovery. Genomics. 2009;93(1):13–16.
3. Mohr S, Liew CC. The peripheral-blood transcriptome: new insights into disease and risk assessment. Trends Mol Med. 2007;13(10):422–32.
4. McLachlan G, Do KA, Ambroise C. Analyzing Microarray Gene Expression Data, vol. 422. Hoboken: Wiley; 2005.
5. Ng SK, McLachlan GJ, Wang K, Nagymanyoki Z, Liu S, Ng SW. Inference on differences between classes using cluster-specific contrasts of mixed effects. Biostatistics. 2015;16(1):98–112.
6. Ben-Dor A, Shamir R, Yakhini Z. Clustering gene expression patterns. J Comput Biol. 2004;6(3–4):281–97.
7. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999;286(5439):531–7.
8. Hartigan JA. Direct clustering of a data matrix. J Am Stat Assoc. 1972;67(337):123–9.
9. Kluger Y, Basri R, Chang JT, Gerstein M. Spectral biclustering of microarray data: coclustering genes and conditions. Genome Res. 2003;13(4):703–16.
10. Ward JH. Hierarchical grouping to optimize an objective function. J Am Stat Assoc. 1963;58(301):236–44.
11. Hartigan JA, Wong MA. A k-means clustering algorithm. Appl Stat. 1979;28(1):100–8.
12. McLachlan GJ, Basford KE. Mixture models: inference and applications to clustering1988.
13. McNicholas PD. Mixture Model-Based Classification. Boca Raton: Chapman & Hall/CRC Press; 2016.
14. Spearman C. The proof and measurement of association between two things. Am J Psychol. 1904;15:72–101.
15. Ghahramani Z, Hinton GE. The EM algorithm for factor analyzers. Technical Report CRG-TR-96-1, University of Toronto. 1997.
16. McLachlan GJ, Peel D. Mixtures of factor analyzers. In: Proceedings of the Seventeenth International Conference on Machine Learning, ICML '00. San Francisco: Morgan Kaufmann Publishers Inc; 2000. p. 599–606.
17. Tipping ME, Bishop CM. Mixtures of probabilistic principal component analyzers. Neural Comput. 1999;11(2):443–82.
18. McNicholas PD, Murphy TB. Parsimonious Gaussian mixture models. Stat Comput. 2008;18:285–96.
19. Pontes B, Giráldez R, Aguilar-Ruiz JS. Biclustering on expression data: A review. J Biomed Inform. 2015;57:163–80.
20. Lazzeroni L, Owen A. Plaid models for gene expression data. Stat Sin. 2000;12:61–86.
21. Sheng Q, Moreau Y, De Moor B. Biclustering microarray data by gibbs sampling. Bioinformatics. 2003;19(suppl 2):196–205.
22. Gu J, Liu JS. Bayesian biclustering of gene expression data. BMC Genomics. 2008;9(1):1–10.
23. Martella F, Alfo M, Vichi M. Biclustering of gene expression data by an extension of mixtures of factor analyzers. Int J Biostat. 2008;4(1). https://www.degruyter.com/downloadpdf/j/ijb.2008.4.1/ijb.2008.4.1.1078/ijb.2008.4.1.1078.pdf.
24. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B Stat Methodol. 1977;39(1):1–38.
25. Meng XL, van Dyk DA. The EM algorithm – an old folk song sung to a fast new tune (with discussion). J R Stat Soc Ser B Stat Methodol. 1997;59:511–67.
26. Woodbury MA. Inverting modified matrices. Memorandum Rep. 1950;42:106.
27. Aitken A. A series formula for the roots of algebraic and transcendental equations. Proc R Soc Edinb. 1926;45(1):14–22.
28. Krishnan T, McLachlan G. The EM Algorithm and Extensions. New York: Wiley; 1997.
29. Bohning D, Dietz E, Schaub R, Schlattmann P, Lindsay BG. The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. Ann Inst Stat Math. 1994;46(2):373–88.
30. McNicholas PD, Murphy TB, McDaid AF, Frost D. Serial and parallel implementations of model-based clustering via parsimonious Gaussian mixture models. Comput Stat Data Anal. 2010;54:711–23.
31. Schwarz G. Estimating the dimension of a model. Ann Stat. 1978;6(2):461–4.
32. Hubert L, Arabie P. Comparing partitions. J Classif. 1985;2:193–218.

33. Biernacki C, Celeux G, Govaert G. Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Trans Pattern Anal Mach Intell. 2000;22(7):719–25.
34. Akaike H. Information theory and an extension of the maximum likelihood principle In: Petrov BN, Csaki F, editors. Proceedings of the 2nd International Symposium on Information Theory. Budapest: Akademiai Kiado; 1973. p. 267–281.
35. Beaudoin MS, Snook LA, Arkell AM, Simpson JA, Holloway GP, Wright DC. Resveratrol supplementation improves white adipose tissue function in a depot-specific manner in Zucker diabetic fatty rats. Am J Physiol Regul Integr Comp Physiol. 2013;305(5):542–51.
36. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30:207–10.
37. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy–analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 2004;20(3):307–15.
38. Carvalho BS, Irizarry RA. A framework for oligonucleotide microarray preprocessing. Bioinformatics. 2010;26(19):2363–367.
39. R Core Team. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing; 2016. R Foundation for Statistical Computing. https://www.R-project.org/.
40. Gentleman RC, Carey VJ, Bates DM, et al. Bioconductor: Open software development for computational biology and bioinformatics. Genome Biol. 2004;5:80.
41. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2008;4:44–57.
42. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37(1):1–13.
43. The Gene Ontology Consortium. Gene Ontology Consortium: going forward. Nucleic Acids Res. 2015;43(D1):1049–56.
44. Issad T, Masson E, Pagesy P. O-GlcNAc modification, insulin signaling and diabetic complications. Diabetes Metab. 2010;36(6, Part 1):423–35.
45. Saltiel AR, Kahn CR. Insulin signalling and the regulation of glucose and lipid metabolism. Nature. 2001;414(6865):799–806.
46. Nawrocki AR, Rajala MW, Tomas E, Pajvani UB, Saha AK, Trumbauer ME, Pang Z, Chen AS, Ruderman NB, Chen H, Rossetti L, Scherer PE. Mice lacking adiponectin show decreased hepatic insulin sensitivity and reduced responsiveness to peroxisome proliferator-activated receptor γ agonists. J Biol Chem. 2006;281(5):2654–660.
47. Laffel L. Ketone bodies: a review of physiology, pathophysiology and application of monitoring to diabetes. Diabetes Metab Res Rev. 1999;15(6):412–26.
48. Yeaman SJ. The 2-oxo acid dehydrogenase complexes: recent advances. Biochem J. 1989;257(3):625–32.
49. Virkamäki A, Ueki K, Kahn CR. Protein–protein interaction in insulin signaling and the molecular mechanisms of insulin resistance. J Clin Investig. 1999;103(7):931–43.
50. Taniguchi CM, Emanuelli B, Kahn CR. Critical nodes in signalling pathways: insights into insulin action. Nat Rev Mol Cell Biol. 2006;7(2):85–96.
51. Kraegen EW, Clark PW, Jenkins AB, Daley EA, Chisholm DJ, Storlien LH. Development of muscle insulin resistance after liver insulin resistance in high-fat-fed rats. Diabetes. 1991;40(11):1397–403.
52. Samuel VT, Petersen KF, Shulman GI. Lipid-induced insulin resistance: unravelling the mechanism. Lancet. 2010;375:2267–277.
53. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A. 2004;101(16):6062–067.
54. Ishii KJ, Akira S. Potential link between the immune system and metabolism of nucleic acids. Curr Opin Immunol. 2008;20(5):524–9.
55. Neefjes J, Jongsma MLM, Paul P, Bakke O. Towards a systems understanding of MHC class I and MHC class II antigen presentation. Nat Rev Immunol. 2011;11(12):823–36.
56. Paludan S, Bowie A. Immune sensing of DNA. Immunity. 2013;38(5):870–80.
57. Parcej D, Tampe R. ABC proteins in antigen translocation and viral inhibition. Nat Chem Biol. 2010;6(8):572–80.
58. Chaussabel D, Pascual V, Banchereau J. Assessing the human immune system through blood transcriptomics. BMC Biol. 2010;8:84–97.
59. Sullivan PF, Fan C, Perou CM. Evaluating the comparability of gene expression in blood and brain. Am J Med Genet B Neuropsychiatr Genet. 2006;141B(3):261–8.
60. Chen J, Chen Z. Extended Bayesian information criteria for model selection with large model spaces. Biometrika. 2008;95(3):759–71.
61. McNicholas PD, Murphy TB. Model-based clustering of microarray expression data via latent Gaussian mixture models. Bioinformatics. 2010;26(21):2705–712.
62. McLachlan GJ, Bean RW, Peel D. A mixture model-based approach to the clustering of microarray expression data. Bioinformatics. 2002;18(3):413–22.