BMC Bioinformatics

RESEARCH

Open Access

A model-based clustering via mixture of hierarchical models with covariate adjustment for detecting differentially expressed genes from paired design

Yixin Zhang[1], Wei Liu[2*] and Weiliang Qiu[3]

*Correspondence:
liuwei@yorku.ca

[1] School of Mathematical Science, University of Science and Technology of China, Hefei, Anhui, China
[2] Department of Mathematics and Statistics, York University, Toronto, ON, Canada
[3] Department of Biostatistics and Programming, Sanofi, Cambridge, MA, USA

Abstract

The causes of many complex human diseases are still largely unknown. Genetics plays an important role in uncovering the molecular mechanisms of complex human diseases. A key step to characterize the genetics of a complex human disease is to unbiasedly identify disease-associated gene transcripts on a whole-genome scale. Confounding factors could cause false positives. Paired design, such as measuring gene expression before and after treatment for the same subject, can reduce the effect of known confounding factors. However, not all known confounding factors can be controlled in a paired/match design. Model-based clustering, such as mixtures of hierarchical models, has been proposed to detect gene transcripts differentially expressed between paired samples. To the best of our knowledge, no model-based gene clustering methods have the capacity to adjust for the effects of covariates yet. In this article, we proposed a novel mixture of hierarchical models with covariate adjustment in identifying differentially expressed transcripts using high-throughput whole-genome data from paired design. Both simulation study and real data analysis show the good performance of the proposed method.

Keywords: Curse of dimensionality, Confounding, EM algorithm, RNAseq

Introduction

Genome-wide differential gene expression analysis is widely used for the elucidation of the molecular mechanisms of complex human diseases. One popular and powerful approach to detect differentially expressed genes is the probe-wise linear regression analysis combined with the control of multiple testing, such as limma [1]. That is, we first perform linear regression for each probe and then adjust p-values for controlling multiple testing. One advantage of this approach is its capacity to adjust for potential confounding factors.

Another approach for detecting differentially expressed genes is the model-based clustering via mixture of Bayesian hierarchical models (MBHM) [2–7], which can borrow

information across genes to cluster genes. Probe clustering based on MBHMs treats gene transcripts as "samples" and samples as "variables". Therefore, transcript clustering based on MBHMs has large number of "samples" and relatively small number of "variables", hence does not have the curse-of dimensionality problem. In addition, unlike transcript-specific tests that have several parameters per transcript, transcript clustering based on MBHMs has only a few hyperparameters per cluster to be estimated and could borrow information across transcripts to estimate model hyperparameters. These approaches generally assume that samples under two groups are obtained independently. [8] proposed a constrained MBHM to identify genetic outcomes measured from paired/matched designs.

Paired design is commonly used in study design for its homogeneous external environment for comparing measurements under different conditions. However, not all known confounding factors can be controlled in a paired/match design. Hence, we might still need to adjust the effects of confounding factors for data from a paired/matched design.

Mixture of regressions or mixture of experts model [9–11] have been proposed in literature to do clustering with capacity to adjust for covariates. To best of our knowledge, this approach does not have constraints on positive, negative, and constant means and has not been applied to detect differentially expressed genes.

In this article, we proposed a novel mixture of hierarchical models with covariate adjustment in identifying differentially expressed transcripts using high-throughput whole genome data from paired design.

Method

We assumed that gene transcripts can be roughly classified into 3 clusters based on their expression levels in subjects after treatment (denoted as condition 1) relative to those before treatment (denoted as condition 2):

1 Transcripts after treatment have higher expression levels than those before treatment, i.e., over-expressed (**OE**) in condition 1;
2 Transcripts after treatment have lower expression levels than those before treatment, i.e., under-expressed (**UE**) in condition 1;
3 Transcripts after treatment have same expression levels than those before treatment, i.e., non-differentially expressed (**NE**) between condition 1 and matched condition 2.

We followed [8] to directly model the marginal distributions of gene transcripts in the 3 clusters. In [8], they proposed a mixture of three-component hierarchical distributions to characterize the within-pair difference of gene expression. We extended their model by incorporating potential confounding factors (such as *Age* and *Sex*) in the mixture of hierarchical models, which might affect the response of gene expression to drug treatment.

Note that this extension is non-trivial, just like multiple linear regression is not just a simple extension to simple linear regression.

We assumed that data have been processed so that the distributions of mRNA expression levels are close to normal distributions. For RNAseq data, we can apply VOOM transformation [12] or countTransformers [13] before applying *eLNNpairedCov*.

A mixture of hierarchical models

For the g^{th} gene transcript, let x_{gl} and y_{gl} denote the expression levels of the l^{th} subject under two different conditions, e.g., before and after treatment, $g = 1, \ldots, G$, $l = 1, \ldots, n$, where G is the number of transcripts and n is the number of subjects (i.e., the number of pairs). Let $d_{gl} = \log_2(y_{gl}) - \log_2(x_{gl})$ be the log2 difference for the g^{th} gene transcript of l^{th} subject. Denote $\mathbf{d}_g = \left(d_{g1}, \ldots, d_{gn}\right)^T$. We assumed that $\mathbf{d}_g$ is conditionally normally distributed given mean vector and covariance matrix. Let $\boldsymbol{W}^T$ be the $n \times (p+1)$ design matrix, where p is the number of covariates. The first column of $\boldsymbol{W}^T$ is the vector of ones, indicating intercept. Let $\boldsymbol{\eta}$ be the $(p+1) \times 1$ vector of coefficients for the intercept and covariate effects. We assume following mixture of three-component hierarchical models:

For gene transcripts over-expressed (OE) in post-treatment samples, we expect that the mean log2 differences are positive. Hence, we assume

$$\begin{aligned}
&\mathbf{d}_g|\left(\boldsymbol{\mu}_g, \tau_g\right) \sim N\left(\boldsymbol{\mu}_g, \tau_g^{-1}\boldsymbol{I}_n\right)\\
&\quad\boldsymbol{\mu}_g|\tau_g \sim N\left(\exp[\boldsymbol{W}^T\boldsymbol{\eta}_1], k_1\tau_g^{-1}\boldsymbol{I}_n\right)\\
&\quad\tau_g \sim \Gamma(\alpha_1, \beta_1)
\end{aligned}$$

where $k_1 > 0, \alpha_1 > 0$ and $\beta_1 > 0$. $\Gamma(\alpha_1, \beta_1)$ denotes the Gamma distribution with shape parameter α_1 and rate parameter β_1. That is, we assume that (1) the mean vectors $\boldsymbol{\mu}_g$, $g = 1, \cdots, G$, given the variance τ_g^{-1} follow a multivariate normal distribution with mean vector $\exp\left[\boldsymbol{W}^T\boldsymbol{\eta}_1\right]$ and covariance matrix $k_1\tau_g^{-1}\boldsymbol{I}_n$; and (2) the variances τ_g^{-1}, $g = 1, \cdots, G$, follow a Gamma distribution with shape parameter α_1 and rate parameter β_1.

Note that the exponential of the intercept $\exp(\eta_{10})$ indicates the mean of log2 difference is positive.

For gene transcripts under-expressed (UE) in post-treatment samples, we expect that the mean log2 differences are negative. Hence, we assume

$$\begin{aligned}
&\mathbf{d}_g|\left(\boldsymbol{\mu}_g, \tau_g\right) \sim N\left(\boldsymbol{\mu}_g, \tau_g^{-1}\boldsymbol{I}_n\right)\\
&\quad\boldsymbol{\mu}_g|\tau_g \sim N\left(-\exp[\boldsymbol{W}^T\boldsymbol{\eta}_2], k_2\tau_g^{-1}\boldsymbol{I}_n\right)\\
&\quad\tau_g \sim \Gamma(\alpha_2, \beta_2)
\end{aligned}$$

where $k_2 > 0, \alpha_2 > 0, \beta_2 > 0$, and $\boldsymbol{W}^T$ is the design matrix.

Note that the negative exponential of the intercept $-\exp(\eta_{20})$ indicates the mean of log2 difference is negative.

For gene transcripts non-differentially expressed (NE) between pre- and post-treatment samples, we expect the mean log2 differences are zero. Hence, we assume

$$\begin{aligned}
&\mathbf{d}_g|\tau_g \sim N\left(\boldsymbol{U}^T\boldsymbol{\theta}_g, \tau_g^{-1}\boldsymbol{I}_n\right)\\
&\quad\boldsymbol{\theta}_g|\tau_g \sim N(\boldsymbol{\eta}_3, k_3\tau_g^{-1}\boldsymbol{I}_p)\\
&\quad\tau_g \sim \Gamma(\alpha_3, \beta_3)
\end{aligned}$$

where $k_3 > 0, \alpha_3 > 0$ and $\beta_3 > 0$. $\boldsymbol{U}^T$ is the design matrix without intercept column. That is, the intercepts are zero. Note that the intercepts indicate mean log2 differences. Hence, $\boldsymbol{\eta}_3$ is a $p \times 1$ vector of coefficients for the covariates.

Note that $\boldsymbol{\theta}_g$ measure effects of confounding factors for NE genes. The true effect of NE genes are zero (i.e., the intercept of $\boldsymbol{U}^T\boldsymbol{\theta}_g$ is zero in the above model).

The hyperparameters α_c and β_c are shape and rate parameters for the Gamma distribution, respectively, $c = 1, 2, 3$. As for k_1, k_2 and k_3, the variation of the mean vector $\boldsymbol{\mu}_g$ should be smaller than that of the observations $\mathbf{d}_g$. So we expect $0 < k_c < 1, c = 1, 2, 3$.

Note that the marginal distribution for each component of the mixture is a multivariate t distribution [14, Section 3.7.6]. However, to model differentially expressed genes, the multivariate t distributions derived from our models have special structure of mean vector and covariance matrix.

For continuous covariates, we require that they are standardized so that they have mean zero and variance one. Standardizing continuous covariates would make $\exp\left(\boldsymbol{W}^T\boldsymbol{\eta}_1\right)$ and $\exp\left(\boldsymbol{W}^T\boldsymbol{\eta}_2\right)$ be numerically finite.

Ideally, we should require $\boldsymbol{\mu}_g > 0$ ($\boldsymbol{\mu}_g < 0$) for all transcripts in cluster 1 (cluster 2). To do so, we can assume a log normal prior distribution for $\boldsymbol{\mu}_g$ in cluster 1, for instance. However, a log normal distribution could not be a conjugate prior for the mean of a normal distribution. It would increase the computational burden if non-conjugate priors were used. Other alternative models can also be used, such as assuming $\boldsymbol{\mu}_g|\eta_{10} = \exp(\eta_{10}) + \boldsymbol{W}^T\boldsymbol{\eta}_1$ and η_{10} follows a normal distribution. However, these models do not have closed-form marginal densities. Hence, they would substantially increase computational burden. Besides, the empirical distribution of the mean log2 difference $\mathbf{d}_g$ of the differentially expressed gene probes has shown a right-skewed pattern, while that of non-differentially expressed genes demonstrates an approximate bell shape (see in Additional file 1: Figures A2-A4). Hence, we require the mean $E(\boldsymbol{\mu}_g) > 0$ ($E(\boldsymbol{\mu}_g) < 0$) for cluster 1 (cluster 2) by assuming $E(\boldsymbol{\mu}_g)$ for cluster 1 (cluster 2) to be $\exp[\boldsymbol{W}^T\boldsymbol{\eta}_1]$ ($-\exp[\boldsymbol{W}^T\boldsymbol{\eta}_2]$).

The proposed mixture models have meaningful biological interpretations for mean structures. In particular, for the **OE** cluster, the intercept $\exp(\eta_{10})$ can be interpreted as the expected average log2 difference of gene transcripts when the value of all the p covariates are zero; the coefficient η_{1i} of covariate i can be interpreted as there exists $\exp(\eta_{1i})$ fold-change associated with the one unit increase in covariate i while the values of the remaining $(p-1)$ covariates are fixed; for the **UE** cluster, the intercept $-\exp(\eta_{20})$ can be interpreted as the expected average log2 difference of gene transcripts when the value of all the p covariates are zero; the coefficient η_{2i} of covariate i can be interpreted as there exists $\exp(\eta_{2i})$ fold-change associated with the one unit increase in covariate i while the values of the remaining $(p-1)$ covariates are fixed; while for the **NE** cluster, the coefficient η_{3i} of covariate i can be interpreted as η_{3i} unit increase of expected log2 difference of gene transcripts associated with the one unit increase in covariate i while the values of the remaining $(p-1)$ covariates are fixed. They also are convenient to get closed-form marginal densities so that we can use Expectation-Maximization (EM) algorithm to estimate hyperparameters, instead of using computational-intensive algorithms, such as Markov chain Monte Carlo (MCMC).

Marginal density functions

Let $f_1(\mathbf{d}_g|\psi)$, $f_2(\mathbf{d}_g|\psi)$, $f_3(\mathbf{d}_g|\psi)$ be the marginal densities of the 3 hierarchical models, and $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)$ be the vector of cluster mixture proportions, where $\psi = \left(\alpha_1, \beta_1, k_1, \boldsymbol{\eta}_1^T, \alpha_2, \beta_2, k_2, \boldsymbol{\eta}_2^T, \alpha_3, \beta_3, k_3, \boldsymbol{\eta}_3^T\right)^T$. Then the marginal density of $\mathbf{d}_g$ is:

$$f(\mathbf{d}_g|\psi) = \pi_1 f_1(\mathbf{d}_g|\psi) + \pi_2 f_2(\mathbf{d}_g|\psi) + \pi_3 f_3(\mathbf{d}_g|\psi).$$

Determining transcript cluster membership

The transcript-cluster membership is determined based on the posterior probabilities, $\zeta_{gc} = Pr(g^{th}$ gene transcript in cluster $c|\boldsymbol{d}_g)$. We can get

$$\zeta_{gc} = \frac{\pi_c f_c(\mathbf{d}_g|\psi)}{\pi_1 f_1(\mathbf{d}_g|\psi) + \pi_2 f_2(\mathbf{d}_g|\psi) + \pi_3 f_3(\mathbf{d}_g|\psi)}, c = 1, 2, 3. \quad (1)$$

We determine a transcript's cluster membership as follows: If the maximum value among $\zeta_{gi}, i = 1, 2, 3$ is ζ_{gc}, then the transcript g belongs to cluster c.

The true values of π_1, π_2, π_3, and ψ are unknown. We use estimated values to determine transcripts' cluster membership.

Parameter estimation via EM algorithm

We used expectation-maximization (EM) algorithm [15] to estimate the model parameters $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)^T$ and ψ.

Let $\boldsymbol{z}_g = (z_{g1}, z_{g2}, z_{g3})$ to be the indicator vector indicating if gene transcript g belongs to a cluster or not. To stablize the estimate of $\boldsymbol{\pi}$ when π_c is very small, we assume that the cluster mixture proportions $\boldsymbol{\pi}$ follows a symmetric Dirichlet $D(\mathbf{b})$ distribution, i.e.,$f(\boldsymbol{\pi}) = \frac{\Gamma(\sum_{c=1}^3 b_c)}{\prod_{c=1}^3 \Gamma(b_c)} \prod_{c=1}^3 \pi_c^{b_c - 1}$. Therefore, the likelihood function for the complete data $(\mathbf{d}, \boldsymbol{z}, \boldsymbol{\pi})$ is

$$\begin{aligned}
L(\psi|\mathbf{d}, \boldsymbol{z}, \boldsymbol{\pi}) &= f(\mathbf{d}, \boldsymbol{z}, \boldsymbol{\pi}|\psi) \\
&= f(\mathbf{d}, \boldsymbol{z}|\boldsymbol{\pi}, \psi) f(\boldsymbol{\pi}|\psi) \\
&= f(\mathbf{d}, \boldsymbol{z}|\boldsymbol{\pi}, \psi) f(\boldsymbol{\pi}) \\
&= \left(\prod_{g=1}^G f(\mathbf{d}_g, \boldsymbol{z}_g|\psi, \boldsymbol{\pi}) \right) Dir(\mathbf{b}) \\
&= \left(\prod_{g=1}^G (\pi_1 f_1(\mathbf{d}_g|\psi))^{z_{g1}} (\pi_2 f_2(\mathbf{d}_g|\psi))^{z_{g2}} (\pi_3 f_3(\mathbf{d}_g|\psi))^{z_{g3}} \right) \\
&\quad \times \frac{\Gamma(\sum_{c=1}^3 b_c)}{\prod_{c=1}^3 \Gamma(b_c)} \prod_{c=1}^3 \pi_c^{b_c - 1}.
\end{aligned}$$

Then the log complete-data likelihood function is:

$$
\begin{aligned}
l(\psi|\mathbf{d},\boldsymbol{z},\boldsymbol{\pi}) &= \sum_{g=1}^{G}((z_{g1}\log f_1(\mathbf{d}_g|\psi) + z_{g2}\log f_2(\mathbf{d}_g|\psi)) + z_{g3}\log f_3(\mathbf{d}_g|\psi)) \\
&+ \sum_{g=1}^{G}(z_{g1}\log\pi_1 + z_{g2}\log\pi_2 + z_{g3}\log\pi_3) \\
&+ \log\left(\frac{\Gamma(\sum_{c=1}^{3} b_c)}{\prod_{c=1}^{3}\Gamma(b_c)}\right) + \sum_{c=1}^{3}(b_c - 1)\log\pi_c.
\end{aligned}
$$

The EM algorithm is used to estimate parameters $\boldsymbol{\pi}$ and ψ. Since $\boldsymbol{z}$ is unknown random vector, we integrate it out from the log complete-data likelihood function. Here, $\boldsymbol{z}_g = (z_{g1}, z_{g2}, z_{g3})$.

$$
\begin{aligned}
\zeta_{g1} &= E(z_{g1}|\mathbf{d}_g,\boldsymbol{\pi},\psi) \\
&= Pr(z_{g1}=1|\mathbf{d}_g,\boldsymbol{\pi},\psi) \\
&= \frac{\pi_1 f_1(\mathbf{d}_g|\psi)}{\pi_1 f_1(\mathbf{d}_g|\psi) + \pi_2 f_2(\mathbf{d}_g|\psi) + \pi_3 f_3(\mathbf{d}_g|\psi)} \\
\zeta_{g2} &= \frac{\pi_2 f_2(\mathbf{d}_g|\psi)}{\pi_1 f_1(\mathbf{d}_g|\psi) + \pi_2 f_2(\mathbf{d}_g|\psi) + \pi_3 f_3(\mathbf{d}_g|\psi)} \\
\zeta_{g3} &= \frac{\pi_3 f_3(\mathbf{d}_g|\psi)}{\pi_1 f_1(\mathbf{d}_g|\psi) + \pi_2 f_2(\mathbf{d}_g|\psi) + \pi_3 f_3(\mathbf{d}_g|\psi)}
\end{aligned}
\tag{2}
$$

E-step. Denote $Q^{(t)}\left(\boldsymbol{\pi},\psi|\mathbf{d},\boldsymbol{z}^{(t)},\boldsymbol{\pi}^{(t)}\right)$ as the expected log complete-data likelihood function at t-th iteration of the EM algorithm, we have

$$
\begin{aligned}
Q^{(t)} &= E_{\boldsymbol{z}}\left[l(\psi|\mathbf{d},\boldsymbol{z},\boldsymbol{\pi})|\boldsymbol{d},\boldsymbol{z}^{(t)},\boldsymbol{\pi}^{(t)}\right] \\
&= \sum_{g=1}^{G}((\zeta_{g1}^{(t)}\log f_1(\mathbf{d}_g|\psi) + \zeta_{g2}^{(t)}\log f_2(\mathbf{d}_g|\psi)) + \zeta_{g3}^{(t)}\log f_3(\mathbf{d}_g|\psi)) \\
&+ \sum_{g=1}^{G}(\zeta_{g1}^{(t)}\log\pi_1 + \zeta_{g2}^{(t)}\log\pi_2 + \zeta_{g3}^{(t)}\log\pi_3) \\
&+ \log\left(\frac{\Gamma(\sum_{c=1}^{3} b_c)}{\prod_{c=1}^{3}\Gamma(b_c)}\right) + \sum_{c=1}^{3}(b_c - 1)\log\pi_c,
\end{aligned}
$$

where

$$
\begin{aligned}
\zeta_{gc}^{(t)} &= E\left(z_{gc}|\mathbf{d}_g,\boldsymbol{\pi}^{(t)},\psi^{(t)}\right) \\
&= \frac{\pi_c^{(t)} f_c(\mathbf{d}_g|\psi^{(t)})}{\pi_1^{(t)} f_1(\mathbf{d}_g|\psi^{(t)}) + \pi_2^{(t)} f_2(\mathbf{d}_g|\psi^{(t)}) + \pi_3^{(t)} f_3(\mathbf{d}_g|\psi^{(t)})},\ c = 1,2,3.
\end{aligned}
\tag{3}
$$

M-step. Maximize $Q^{(t)}\left(\boldsymbol{\pi},\psi|\mathbf{d},\boldsymbol{z}^{(t)},\boldsymbol{\pi}^{(t)}\right)$ to find the optimal values of $\boldsymbol{\pi}$ and ψ, and use these optimal values as estimates for the parameters $\boldsymbol{\pi}$ and ψ.

To maximize $Q^{(t)}\left(\boldsymbol{\pi},\psi|\mathbf{d},\boldsymbol{z}^{(t)},\boldsymbol{\pi}^{(t)}\right)$, we use the "L-BFGS-B" method developed by Byrd et al. (1995) [16], which utilizes the first partial derivatives of $Q^{(t)}\left(\boldsymbol{\pi},\psi|\mathbf{d},\boldsymbol{z}^{(t)},\boldsymbol{\pi}^{(t)}\right)$ and allows box constraints, that is each variable can be given a lower and/or upper bound.

Simulated annealing modification

EM algorithm may be trapped in a local maximum since it is strictly ascending. As introduced by Celeux and Govaert (1992) [17], simulated annealing (SA) is widely used to help EM algorithm escape from local maximum by adding randomness with a stochastic step. Specifically, the conditional expectation in (2) is modified in a SA algorithm as follows

$$\tilde{\zeta}_{gc}^{(t)} = \frac{\left[\pi_c^{(t)} f_c(\mathbf{d}_g|\psi^{(t)})\right]^{1/m^{(t)}}}{\sum_{c=1}^{3}\left[\pi_c^{(t)} f_c(\mathbf{d}_g|\psi^{(t)})\right]^{1/m^{(t)}}},\ c = 1, 2, 3. \tag{4}$$

where m is the temperature used to control the randomness. Usually, the temperature m starts with a relatively high value since larger m leads to larger randomness. At iteration t, the temperature is updated by $m^{(t+1)} = r \times m^{(t)}$ with the cooling rate r controls the speed of reduction. As suggested in [18, 19], we use $m^{(0)} = 2$ and $r = 0.9$.

We denoted *eLNNpairedCov* as the proposed method using the traditional EM algorithm to obtain parameter estimates and denoted *eLNNpairedCov.SEM* as the proposed method using the EM with SA-modification to obtain parameter estimates.

We stop the expectation-maximization iterations based on a proportional change, i.e. if the maximum of the absolute value of the differences of model parameter estimates between current iteration and previous iteration over the absolute value of the previous iteration estimates is smaller than a small constant (e.g. 1.0×10^{-3}).

More details about the EM algorithm are shown in Supplementary Document [see Additional file 1].

A real data study

We used the dataset GSE24742 [20], which can be downloaded from the Gene Expression Omnibus [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE24742], to evaluate the performance of the proposed model-based clustering methods (denoted as *eLNNpairedCov* and *eLNNpairedCov.SEM*).

The dataset is from a study that investigated the gene expression before and after administrating rituximab, a drug for treating anti-TNF resistant rheumatoid arthritis (RA). There are 12 subjects, each having 2 samples (one sample is before treatment and the other is after treatment). Age and sex are also available. Expression levels of 54,675 gene probes were measured for each of the 24 samples by using Affymetrix HUman Genome U133 Plus 2.0 array. The dataset has been preprocessed by the dataset contributor. We further kept only 43,505 gene probes in the autosomal chromosomes (i.e., chromosomes 1 to 22). We then performed log2 transformation for gene expression levels. We next obtained the within-subject difference of the log2 transformed expression levels (log2 expression after-treatment minus log2 expression before-treatment). By examining the histogram (Figure A1) [see Additional file 1] of the estimated standard deviations of log2 differences of within-subject gene expression for the 43,505 gene probes, we found a bimodal distribution. Based on Figure A1 [see Additional file 1], where the histogram of estimated standard deviations exhibits two modes, we choose to exclude gene probes with standard deviation < 1 corresponding to the first mode. It is a common practice to

remove genes with low variation [21–23]. Finally, 23,948 gene probes kept in the downstream analysis.

A simulation study

We performed a simulation study to compare the performance of the proposed methods *eLNNpairedCov*, *eLNNpairedCov.SEM* with transcript-wise test *limma* and Li et al.'s [8] method (denoted as *eLNNpaired*). *eLNNpairedCov*, *eLNNpairedCov.SEM* and *limma* adjust covariate effects, while *eLNNpaired* does not. For *eLNNpaired*, we first regress out covariates effect for each gene to make a fair comparison between *eLNNpaired* and other methods.

The *limma* approach first performs an empirical-Bayes-based linear regression for each transcript. In this linear regression, the within-subject log2 difference of transcript expression is the outcome and intercept indicating if the transcript is over-expressed (intercept>0), under-expressed (intercept<0), or non-differentially expressed (intercept $= 0$), adjusting for potential confounding factors. A transcript is claimed as OE if its intercept estimate is positive and corresponding FDR-adjusted p-value < 0.05, where FDR stands for false discovery rate. A transcript is claimed as UE if its intercept estimate is negative and corresponding FDR-adjusted p-value < 0.05. Other transcripts are claimed as NE.

The parameter values ($\boldsymbol{\pi}$, ψ, and proportion of women) in the simulation study are based on the estimates via *eLNNpairedCov.SEM* from the analysis of the pre-processed real dataset GSE24742 described in Subsection "*A real data study*".

In this simulation study, we considered two sets with different covariate coefficients for differentially expressed genes clusters. In the first set (Set 1), parameter values are the estimates of parameters based on the *eLNNpairedCov.SEM* method from real dataset. That is, $\pi_1 = 0.00246$, $\pi_2 = 0.01470$, $\pi_3 = 0.98284$, $\alpha_1 = 3.53$, $\beta_1 = 3.45$, $k_1 = 0.26$, $\eta_{10} = 0.18$, $\eta_{11} = 0.00$, $\eta_{12} = -1.05$, $\alpha_2 = 3.53$, $\beta_2 = 3.45$, $k_2 = 0.26$, $\eta_{20} = 0.18$, $\eta_{21} = 0.00$, $\eta_{22} = -1.05$, $\alpha_3 = 2.86$, $\beta_3 = 2.20$, $k_3 = 0.72$, $\eta_{31} = -0.01$, $\eta_{32} = 0.00$. In the second set (Set 2), we set η_{10}=η_{20}=0.08 instead of 0.18. For each set, we considered two scenarios. In the first scenario (Scenario1), the number of subjects is equal to 30. In the second scenario (Scenario2), the number of subjects is equal to 100.

For each scenario, we generated 100 datasets. Each simulated dataset contains $G = 20,000$ gene transcripts. There are two covariates: standardized age (denoted as *Age.s*) and *Sex*. *Age.s* follows normal distribution with mean 0 and standard deviation 1. Seventy five percent (75%) of subjects are women.

Evaluation criteria

Two agreement indices and two error rates are used to compare the predicted cluster membership and true cluster membership of all genes. The two agreement indices are accuracy (i.e., proportion of predicted cluster membership equal to the true cluster membership) and Jaccard index [24]. For perfect agreement, these indices have a value of one. If an index takes a value close to zero, then the agreement between the true transcript cluster membership and the estimated transcript cluster membership is likely due to chance. The two error rates are false positive rate (FPR) and false negative rate (FNR). FPR is the percentage of detected DE transcripts among truly NE transcripts. FNR is the

Table 1 Parameter estimates of OE, UE and NE clusters from *eLNNpairedCov* and *eLNNpairedCov.SEM*

OE		UE		NE	
β_1	3.445543	β_2	3.445543	β_3	3.445543
k_1	0.264565	k_2	0.264565	k_3	0.264565
η_{10}	0.176007	η_{20}	0.176007		
η_{11}	−0.000609	η_{21}	−0.000609	η_{31}	−0.013796
η_{12}	−1.051257	η_{22}	−1.051257	η_{32}	−0.000017

percentage of detected NE transcripts among truly DE transcripts. We also examined the user time and number of EM iterations for running each simulated dataset.

Results

Results of the real data analysis

For the real dataset, we adjusted standardized age and sex for *eLNNpairedCov*, *eLNNpairedCov.SEM*, and *limma*. We standardized age so that it has mean zero and variance one. For each transcript, we also scaled its expression across subjects so that its variance is equal one. For *eLNNpaired*, we first regressed out the effect of standardized age and sex for each transcript.

The estimates of parameters in our model are listed in Table 1. Note that the proposed *eLNNpairedCov* and *eLNNpairedCov.SEM* have the same estimates for the parameters in these three clusters, except for the proportions of three clusters. The proportions of OE and UE estimated by *eLNNpairedCov* method are 0.0376% and 0.346%, respectively. The proportions of OE and UE estimated by *eLNNpairedCov.SEM* method are 0.246% and 1.47%, respectively.

For the OE cluster, $\exp(\eta_{10}) = \exp(0.176007) = 1.192$ can be interpreted as the expected log2 difference for a male subject ($sex = 0$) whose age is equal to mean age ($age = 0$ is the mean-centered age); $\eta_{11} = -0.000609$ indicates that one-unit increase in *age* leads to $\exp(-0.000609) = 0.999$ fold-changes in expected log2 difference, while $\eta_{12} = -1.051257$ indicates that there is $\exp(-1.051257) = 0.349$ fold-changes between male subjects and female subjects in expected log2 difference if they are at the same *age*. For the UE cluster, $\eta_{20} = 0.176007$ can be interpreted as the expected log2 difference for a male subject ($sex = 0$) whose age is equal to mean age ($age = 0$ is the mean-centered age) is $-\exp(0.176007) = -1.192$; $\eta_{21} = -0.000609$ indicates that one-unit increase in *age* leads to $\exp(-0.000609) = 0.999$ fold-changes in expected log2 difference, while $\eta_{22} = -1.051257$ indicates that there is $\exp(-1.051257) = 0.349$ fold-changes between male subjects and female subjects in expected log2 difference if they are at the same *age*. For the NE cluster, $\eta_{31} = -0.013796$ indicates that one-unit increase in *age* leads to 0.01379 decreases in expected log2 difference, and $\eta_{32} = -0.000017$ indicates that there is 0.000017 decrease from female subjects to male subjects in the expected log2 difference if they are at the same *age*.

The number of differentially expressed genes detected by each method is listed in Table 2.

The *limma* method detected 6 under-expressed gene transcripts (Figure 1 and Table S1), while *eLNNpaired* did not find any positive signals (i.e., $\hat{\pi}_3 = 1$). The

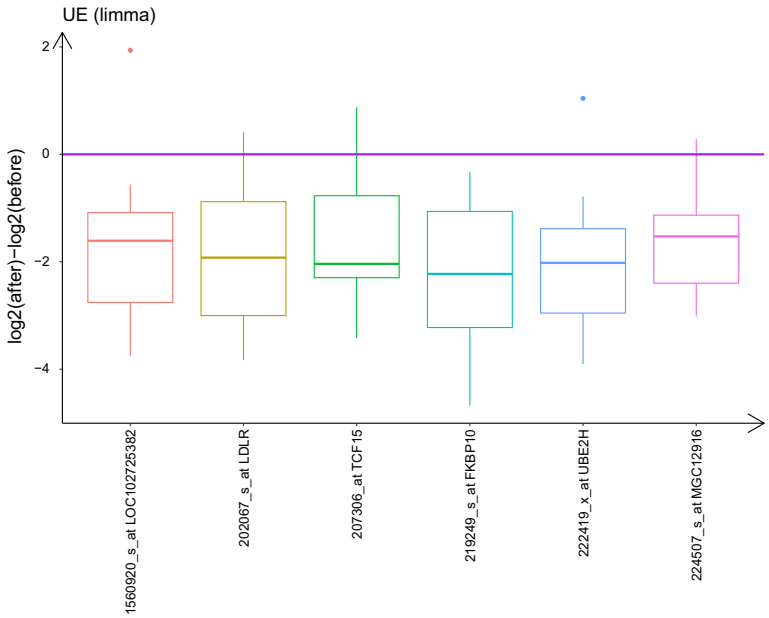

Fig. 1 Parallel boxplots of log2 within-subject difference of gene expression for 6 UE transcripts detected by *limma* for pre-processed GSE24742 dataset. Red horizontal line indicates log2 difference equal to zero

proposed methods *eLNNpairedCov* and *eLNNpairedCov.SEM* detected 55 OE transcripts (Table S2) and 59 OE transcripts (Table S3), respectively (Upper two panels of Fig. 2) and 355 UE transcripts (Table S4) and 352 UE transcripts (Table S5), respectively (Lower two panels of Figure 2). The 6 UE transcripts detected by *limma* is also selected as UE transcripts by *eLNNpairedCov* and *eLNNpairedCov.SEM*. Note that the 55 OE genes detected by *eLNNpairedCov* are also detected by *eLNNpariedCov.SEM*. The 352 UE genes detected by *eLNNpairedCov.SEM* are also detected by *eLNNpariedCov*.

It is assuring that several genes corresponding to the DE transcripts identified by *eLNNpairedCov* and *eLNNpairedCov.SEM* have been associated to rheumatoid arthritis (RA) in literature. For example, Humby et al. (2019) [25] reported that genes *ZNF365* (OE), *IL36RN* (OE), *MRVI1-AS1* (OE), *WFDC6* (UE), *UBE2H* (UE), are associated with RA.

We performed pathway enrichment analysis through the use of IPA (QIAGEN Inc., https://www.qiagenbioinformatics.com/products/ingenuitypathway-analysis) for 352 UE and 55 OE genes identified by *eLNNpairedCov.SEM*. The top enriched canonical pathways are shown in Tables 3 and 4. Evidence in literature shows that these pathways are relevant to RA. S100 protein family plays an important role in rheumatoid arthritis ([26]). Literature shows consistent crucial role of the PD-1/PD-L pathway in the pathogenesis of rheumatic diseases ([27, 28]). It has been shown that RA can lead to lung tissue damage, resulting in pulmonary fibrosis ([29]). Macrophage is a key player in the pathogenesis of autoimmune diseases, such as RA ([30]). RA and osteoarthritis (OA) are two common arthritis with different pathogenesis ([31]). It is interesting to see Osteoarthritis pathway is a significantly enriched pathway for UE genes. It is consistent with literature that similar focal and systemic alterations exist in RA and OA [32].

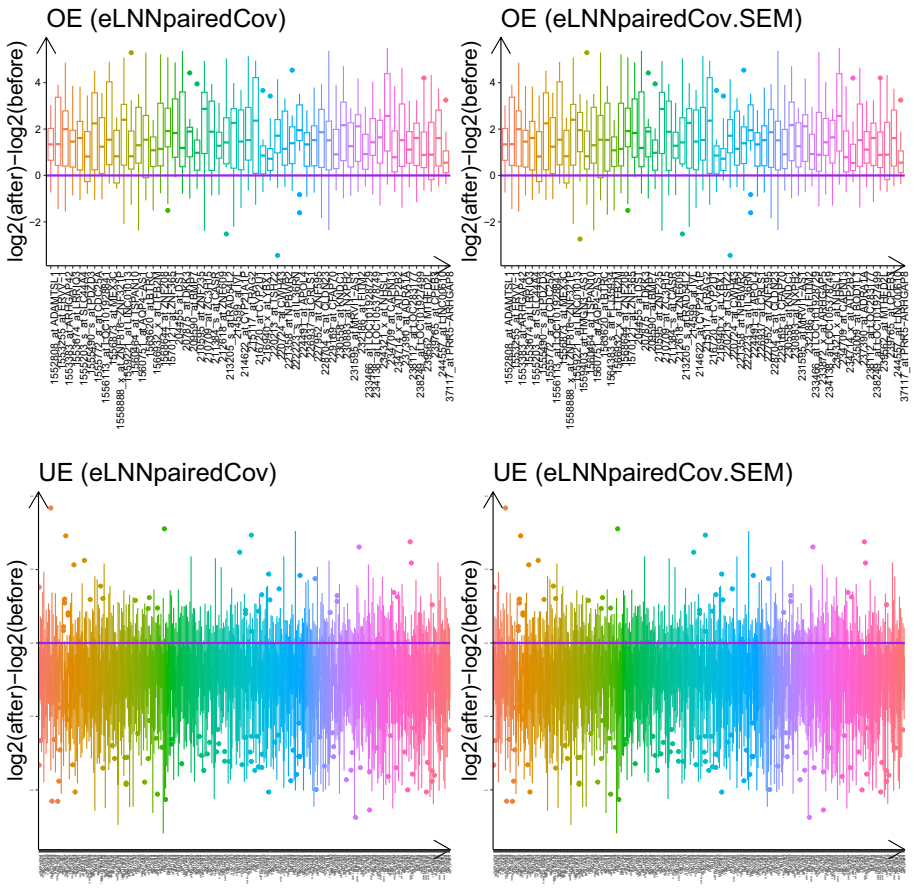

Fig. 2 Parallel boxplots of log2 within-subject difference of gene expression for differentially expressed transcripts detected by *eLNNpairedCov* and *eLNNpairedCov.SEM* for pre-processed GSE24742 dataset. Upper two panels: 55 OE transcripts and 59 OE transcripts, respectively; Lower two panels: 355 UE transcripts and 352 UE transcripts, respectively. Red horizontal lines indicate log2 difference equal to zero

Table 2 Number of Differentially expressed genes detected by *limma*, *eLNNpaired*, *eLNNpairedCov* and *eLNNpairedCov.SEM* in GSE24742

	limma	eLNNpaired	eLNNpairedCov	eLNNpairedCov.SEM
OE	0	0	55	59
UE	6	0	355	352

Table 3 Top canonical pathways for 352 UE genes by *eLNNpairedCov.SEM*

Name	p-value
S100 Family Signaling Pathway	$2.97E-06$
PD-1, PD-L1 cancer immunotherapy pathway	$7.54E-05$
Pulmonary Fibrosis Idiopathic Signaling pathway	$3.45E-04$
Phagosome Formation	$7.56E-04$
Osteoarthritis Pathway	$1.04E-03$

Table 4 Top canonical pathways for 55 OE genes by *eLNNpairedCov.SEM*

Name	p-value
Ribonucleotide Reductase Signaling Pathway	$5.34E-03$
Leukocyte Extravasation Signaling	$7.57E-03$
Cell Cycle: G1/S Checkpoint Regulation	$8.85E-03$
Tetrahydrofolate Salvage from 5,10- methenyltetrahydrofolate	$1.04E-02$
Role of Osteoblasts, Osteoclasts and Chondrocytes in Rheumatoid Arthritis	$1.19E-02$

Ribonucleotide Reductase (RNR) is the enzyme providing the precursors needed for both synthesis and repair of DNA, which could be a potential drug for RA ([33, 34]). Leukocyte extravasation through the endothelial barrier is important in the pathogenesis of RA ([35]). It has been shown that the limb bud and heart development (LBH) gene is a key dysregulated gene in RA and other autoimmune diseases and there are some evidence showing LBH could modulate the cell cycle [36]. Osteoblasts, osteoclasts and chondrocytes play importan roles in Rheumatoid Arthritis ([37–39]). We did not find literature linking Tetrahydrofolate Salvage from 5,10- methenyltetrahydrofolate to RA yet, indicating this enrichment might be novel.

Results of the simulation study

For Scenario 1 ($n = 30$), the jittered scatter plots of the performance indices versus methods are shown in Fig. 3 (Set 1) and Fig. 5 (Set 2) and the jittered scatter plots of the difference of the performance indices versus methods are shown in Fig. 4 (Set 1) and Figure 6 (Set 2).

The differences of performance indices are between *eLNNpairedCov.SEM* and the other three methods (*limma*, *eLNNpaired* and *eLNNpairedCov*). A positive difference indicates that the performance indices of the other method is larger than that of *eLNNpairedCov.SEM*. A negative difference indicates that the performance indices of the other method is smaller than that of *eLNNpairedCov.SEM*.

The upper panel of Figs. 3, 4, 5 and 6 show that both the *eLNNpairedCov* and *eLNNpairedCov.SEM* have higher agreement indices (Jaccard and accuracy) than *limma*, which in turn have higher agreement indices than *eLNNpaired*.

The middle panel of Figures 3-6 show that the proposed *eLNNpairedCov* and *eLNNpairedCov.SEM* methods have similar performance, They have lower FPR than *limma*, while *eLNNpaired* has an exceedingly low FPR (close to 0). The middle panel also show that *eLNNpairedCov*, *eLNNpairedCov.SEM* have smaller FNR than *limma*, while *eLNNpaired* has an exceedingly high FNR (close to 1). The extreme values in FPR and FNR of *eLNNpaired* can be attributed to the fact that it did not detect any differentially expressed genes in this case.

Additionally, Figs. 3, 4, 5 and 6 also show that compared with the performances of these methods in Set 1 ($\eta_{101} = \eta_{20} = 0.18$), those in Set 2 ($\eta_{101} = \eta_{20} = 0.08$) have lower agreement indices and higher error rates except for *eLNNpaired*, which fails to detect any differentially expressed genes in both Set 1 and Set 2.

The bottom panel of Figs. 3 and 5 show that *limma* runs very fast, while *eLNNpaired*, *eLNNpairedCov* and *eLNNpairedCov.SEM* run in reasonable time (i.e., less than 30 s

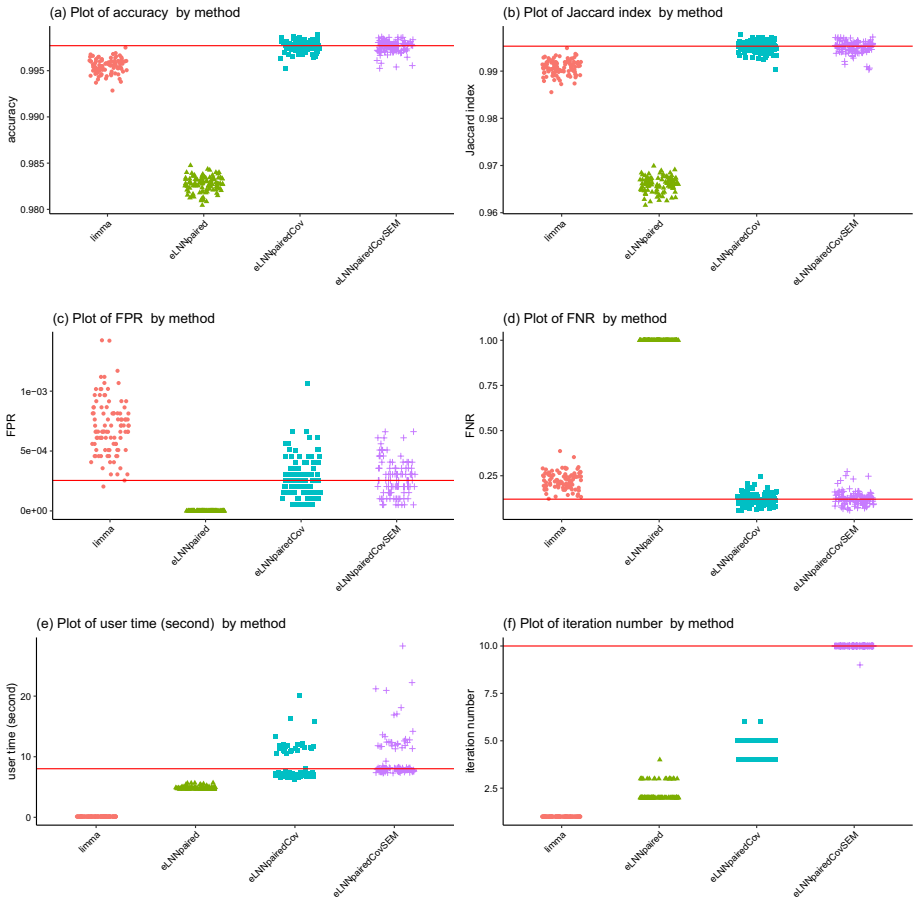

Fig. 3 Jittered scatter plots of performance indices versus method for Set 1, Scenario 1 (number of pairs= 30). Red solid horizontal lines indicate the median performance indices of *eLNNpairedCov.SEM*

per dataset that has $G = 20,000$ genes and $n = 30$ subjects). On average *eLNNpairedCov* and *eLNNpairedCov.SEM* spend a little more time than *eLNNpaired*. The bottom panel of Fig. 3 and 5 also show that *eLNNpaired* uses less than 5 EM iterations, while *eLNNpairedCov* and *eLNNpairedCov.SEM* tend to use more EM iterations. In particular, *eLNNpairedCov.SEM* uses 10 EM iterations, which is the maximum number of iterations we set to save computing time. Note that the EM iteration number for *limma* is set to be one, which does not use EM algorithm to obtain parameter estimates.

The simulation results for Scenario 2 ($n = 100$) are shown in Figures A5-A8 [see Additional file 1], which have similar patterns to those for Scenario 1 ($n = 30$), except that both *eLNNpairedCov* and *eLNNpairedCov.SEM* have smaller FPR which are close to 0. Note that *eLNNpairedCov*,*eLNNpairedCov.SEM* and *limma* have small FNR (close to 0), while *eLNNpaired* still has huge FNR (close to 1).

Discussion and conclusion

In this article, we proposed a novel model-based clustering approach to detect differential expressed transcripts between samples before treatment and samples after treatment, with the capacity to adjust for potential confounding factors. This is novel in that

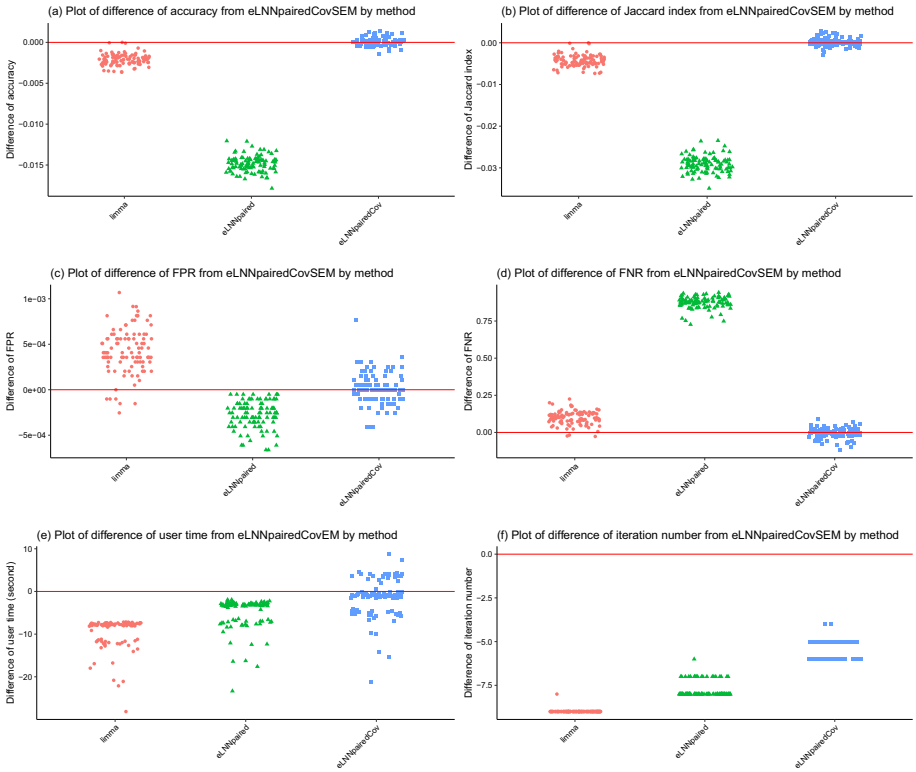

Fig. 4 Jittered scatter plots of difference of performance indices versus method for Set 1, Scenario 1 (number of pairs= 30). Red solid horizontal lines indicate y-axis equal to zero

to the best of our knowledge, all existing model-based gene clustering methods do not yet have the capacity to adjust for covariates.

The proposed approach is different from transcript-wise test followed by multiplicity adjustment in that it does not involve hypothesis testing. Hence, no multiplicity adjustment is needed. The simulation study showed that if the difference of gene expression between samples before treatment and samples after treatment follows the mixture of hierarchical models in Subsection "*A mixture of hierarchical models*", then the proposed method can outperform *limma*, which is a fast and powerful transcript-wise test method. The real data analysis also showed the proposed method *eLNNpairedCov* can detect more differentially expressed gene transcripts, which include the transcripts detected by *limma*.

Although we classify genes to three distinct clusters, the transitions between these clusters could be smooth. This would be reflected by a gene's posterior probability that might be large in two of three clusters, e.g., 0.49 for cluster 1, 0.01 for cluster 2, and 0.5 for cluster 3. On the other hand, expression changes could be split up into more than 3 clusters, e.g., groups behaving differently. In this article, we are only interested in identifying three clusters of genes: over-expressed in condition 1, under-expressed in condition 1, and non-differentially expressed.

There are other model-based clustering methods in literature, such as [40]. However, they were not designed to detect differentially expressed genes. For example, we can set the number K of clusters as 3 for their model. However, there is no constraints that the

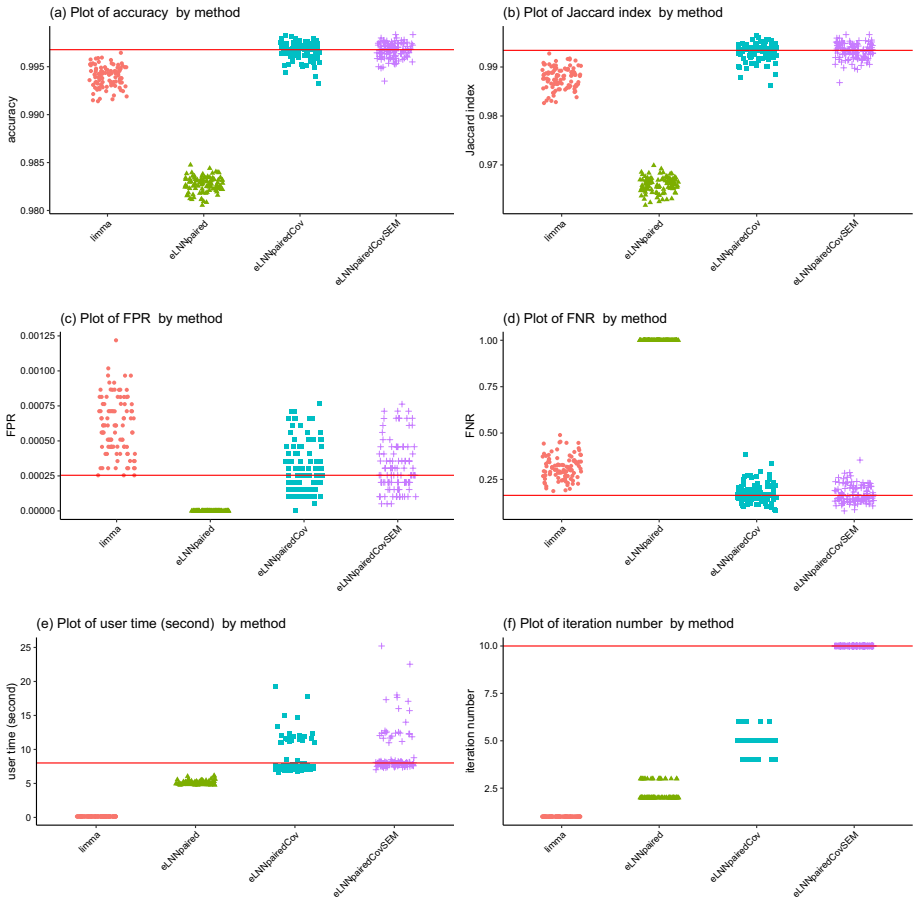

Fig. 5 Jittered scatter plots of performance indices versus method for Set 2, Scenario 1 (number of pairs= 30). Red solid horizontal lines indicate the median performance indices of *eLNNpairedCov.SEM*

intercepts for the three clusters have to be positive, negative, and zero. That is, the three clusters identified might not correspond to over-expressed, under-expressed, and non-differentially expressed genes.

It is well-known in literature that EM algorithm might stuck at local optimal solution. In this article, we used EM with SA-modification to help escape from local optimal solutions. In future, we plan to try the hybrid algorithm of the DPSO (Discrete Particle Swarm Optimization) and the EM approach to improve the global search performance [41].

In our models, the three gene groups allow to have different coefficients of covariates. In future, we could test if these coefficients are same or not. If no significant difference, we could use a model assuming equal coefficients.

RNAseq and single-cell RNAseq data are cutting-edge tools to investigate molecular mechanisms of complex human diseases. However, it is quite challenging to analyze these count data with inflated zero counts. In future, we will evaluate if eLNNpairedCov can be used to analyze single-cell RNAseq data by first transforming counts to continuous scale (e.g., via VOOM [12] or countTransformers [13]) and then to apply eLNNpairedCov to the transformed data.

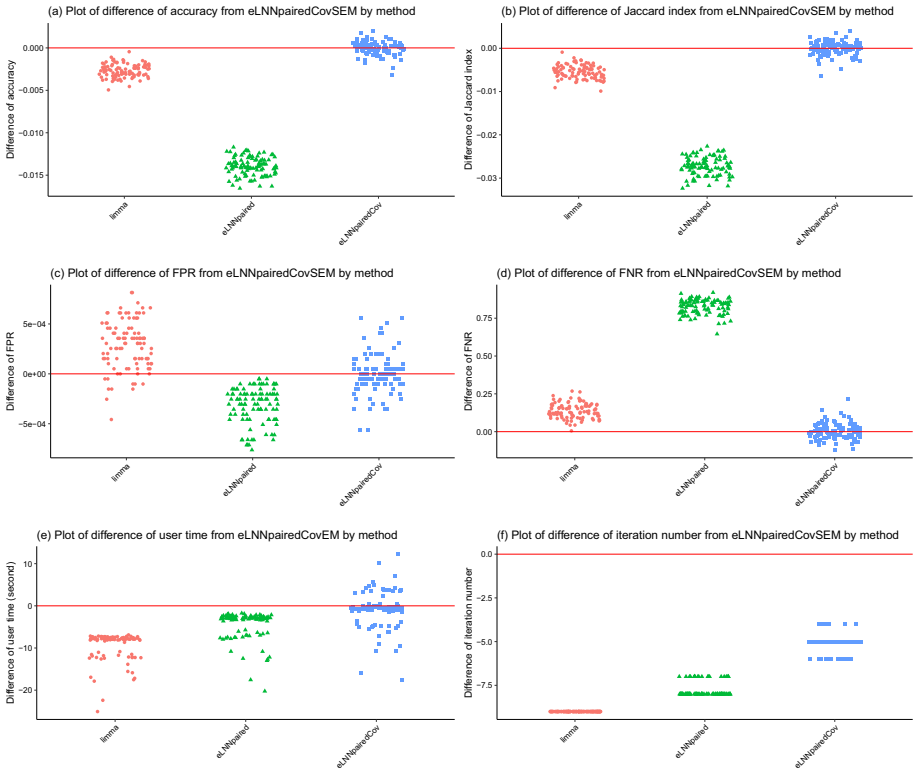

Fig. 6 Jittered scatter plots of difference of performance indices versus method for Set 2, Scenario 1 (number of pairs= 30). Red solid horizontal lines indicate y-axis equal to zero

We implemented the proposed methods to an R package *eLNNpairedCov*, which will be freely available to researchers.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-023-05556-x.

Additional file 1. Supplementary Document.

Additional file 2: Table S1.Gene list of 6 UE transcripts detected by limma.

Additional file 3: Table S2.Gene list of 55 OE transcripts detected by eLNNpairedCov.

Additional file 4: Table S3.Gene list of 59 OE transcripts detected by eLNNpairedCov.SEM.

Additional file 5: Table S4.Gene list of 355 UE transcripts detected by eLNNpairedCov.

Additional file 6: Table S5.Gene list of 352 UE transcripts detected by eLNNpairedCov.SEM.

Acknowledgements
Not applicable

Author contributions
Conceptualization,W.L. and W.Q.; methodology, Y.Z., W.L, and W.Q.; software, Y.Z. and W.Q.; validation, Y.Z., W.L. and W.Q.; formal analysis, Y.Z. and W.Q.; investigation, W.L.; resources, W.L.; data curation, Y.Z.; writing-original draft preparation, Y.Z.; writing-review and editing, W.L. and W.Q.; visualization, Y.Z. and W.Q.; supervision, W.L. and W.Q.; project administration, W.L.; funding acquisition, W.L. All authors have read and agreed to the published version of the manuscript.

Funding
This article was supported by Canada Natural Sciences and Engineering Research Council (NSERC) grants 198662. W.Q. is a Sanofi employee and may hold shares and/or stock options in the company.

Availability of data material
The real dataset analyzed during the current study are available in the Gene Expression Omnibus (GEO) repository, [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE24742].

Declarations

Ethics approval and consent to participate
Not Applicable.

Consent for publication
Not Applicable.

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

Received: 7 April 2023 Accepted: 30 October 2023
Published online: 08 November 2023

References

1. Smyth GK. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology 2004;**3**(1)
2. Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui K-W. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. Journal of Computational Biology. 2001;8(1):37–52.
3. Baldi P, Long AD. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. Bioinformatics. 2001;17(6):509–19.
4. Kendziorski C, Newton M, Lan H, Gould M. On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. Statistics in Medicine. 2003;22(24):3899–914.
5. Gottardo R, Pannucci JA, Kuske CR, Brettin T. Statistical analysis of microarray data: a Bayesian approach. Biostatistics. 2003;4(4):597–620.
6. Lo K, Gottardo R. Flexible empirical Bayes models for differential gene expression. Bioinformatics. 2007;23(3):328–35.
7. Zuyderduyn SD. Statistical analysis and significance testing of serial analysis of gene expression data using a Poisson mixture model. BMC Bioinformatics 2007;**8** . Article number: 283
8. Li Y, Morrow J, Raby B, Tantisira K, Weiss ST, Huang W, Qiu W. Detecting disease-associated genomic outcomes using constrained mixture of Bayesian hierarchical models for paired data. Plos One. 2017;12(3):0174602.
9. Jacobs RA, Jordan MI, Nowlan SJ, Hinton GE. Adaptive mixtures of local experts. Neural Computation. 1991;3(1):79–87.
10. Gormley IC, Frühwirth-Schnatter S. Mixture of experts models. In: Handbook of Mixture Analysis, pp. 271–307. Chapman and Hall/CRC, Boca Raton, FL, USA 2019.
11. Courbariaux M, De Santiago K, Dalmasso C, Danjou F, Bekadar S, Corvol J-C, Martinez M, Szafranski M, Ambroise C. A sparse mixture-of-experts model with screening of genetic associations to guide disease subtyping. Frontiers in Genetics. 2022;13: 859462.
12. Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology. 2014;15(2):1–17.
13. Zhang Z, Yu D, Seo M, Hersh CP, Weiss ST, Qiu W. Novel data transformations for RNA-seq differential expression analysis. Scientific Reports. 2019;9(1):4820.
14. Lenk P. Bayesian inference and Markov chain Monte Carlo. https://webuser.bus.umich.edu/plenk/Bam2%20Short.pdf 2001.
15. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: series B (methodological). 1977;39(1):1–22.
16. Byrd RH, Lu P, Nocedal J, Zhu C. A limited memory algorithm for bound constrained optimization. SIAM Journal on Scientific Computing. 1995;16(5):1190–208.
17. Celeux G, Govaert G. A classification EM algorithm for clustering and two stochastic versions. Computational statistics & Data analysis. 1992;14(3):315–32.
18. Van Laarhoven PJ, Aarts EH. Simulated annealing. Simulated Annealing: Theory and Applications, pp. 7–15. Springer, Dordrecht, Ho11and 1987.
19. Qiao Z, Barnes E, Tringe S, Schachtman DP, Liu P. Poisson hurdle model-based method for clustering microbiome features. Bioinformatics. 2023;39(1):782.
20. Gutierrez-Roelens I LB. Effects of Rituximab on global gene expression profiles in the RA synovium. *NCBI*https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE24742 *2010*.
21. Calza S, Raffelsberger W, Ploner A, Sahel J, Leveillard T, Pawitan Y. Filtering genes to improve sensitivity in oligonucleotide microarray data analysis. Nucleic Acids Research. 2007;35(16): e102.
22. Hackstadt AJ, Hess AM. Filtering for increased power for microarray data analysis. BMC Bioinformatics 2009;**10**(1)
23. Bourgon R, Gentleman R, Huber W. Independent filtering increases detection power for high-throughput experiments. Proceedings of the National Academy of Sciences. 2010;107(21):9546–51.
24. Milligan GW, Cooper MC. A study of the comparability of external criteria for hierarchical cluster analysis. Multivariate Behavioral Research. 1986;21(4):441–58.
25. Humby F, Lewis M, Ramamoorthi N, Hackney JA, Barnes MR, Bombardieri M, Setiadi AF, Kelly S, Bene F, DiCicco M, et al. Synovial cellular and molecular signatures stratify clinical response to csDMARD therapy and predict radiographic progression in early rheumatoid arthritis patients. Annals of the Rheumatic Diseases. 2019;78(6):761–72.
26. Wu Y-Y, Li X-F, Wu S, Niu X-N, Yin S-Q, Huang C, Li J: Role of the S100 protein family in rheumatoid arthritis. Arthritis Research & Therapy 2022;**24** . Article number: 35

27. Zhang S, Wang L, Li M, Zhang F, Zeng X. The PD-1/PD-L pathway in rheumatic diseases. Journal of the Formosan Medical Association 120(1, Part 1), 2021;48–59
28. Canavan M, Floudas A, Veale DJ, Fearon U. The PD-1: PD-L1 axis in inflammatory arthritis. BMC Rheumatology. 2021;5(1):1–10.
29. Lee H, Lee S-I, Kim H-O. Recent advances in basic and clinical aspects of rheumatoid arthritis-associated interstitial lung diseases. Journal of Rheumatic Diseases. 2022;29(2):61–70.
30. Yang S, Zhao M, Jia S. Macrophage: key player in the pathogenesis of autoimmune diseases. Frontiers in Immunology. 2023;14:1080310.
31. Huang H, Dong X, Mao K, Pan W, Nie B, Jiang L. Identification of key candidate genes and pathways in rheumatoid arthritis and osteoarthritis by integrated bioinformatical analysis. Frontiers in Genetics. 2023;14:1083615.
32. Malemud CJ, Schulte ME. Is there a final common pathway for arthritis? International Journal of Clinical Rheumatology. 2008;3(3):253–68.
33. Wang X, Wang X, Sun J, Fu S. An enhanced RRM2 siRNA delivery to rheumatoid arthritis fibroblast-like synoviocytes through a liposome-protamine-DNA-siRNA complex with cell permeable peptides. International Journal of Molecular Medicine. 2018;42(5):2393–402.
34. Huang J-B, Chen Z-R, Yang S-L, Hong F-F. Nitric oxide synthases in rheumatoid arthritis. Molecules. 2023;28(11):4414.
35. Szekanecz Z, Koch AE. Endothelial cells and immune cell migration. Arthritis Research & Therapy 2000;**2** . Article number: 368
36. Matsuda S, Hammaker D, Topolewski K, Briegel KJ, Boyle DL, Dowdy S, Wang W, Firestein GS. Regulation of the cell cycle and inflammatory arthritis by the transcription cofactor LBH gene. The Journal of Immunology. 2017;199(7):2316–22.
37. Berardi S, Corrado A, Maruotti N, Cici D, Cantatore F. Osteoblast role in the pathogenesis of rheumatoid arthritis. Molecular Biology Reports. 2021;48(3):2843–52.
38. Jeong W-J, Kim H-J. Osteoclasts: crucial in rheumatoid arthritis. Journal of Rheumatic Diseases. 2016;23(3):141–7.
39. Tseng C-C, Chen Y-J, Chang W-A, Tsai W-C, Ou T-T, Wu C-C, Sung W-Y, Yen J-H, Kuo P-L. Dual role of chondrocytes in rheumatoid arthritis: the chicken and the egg. International Journal of Molecular Sciences. 2020;21(3):1071.
40. Grün B, Leisch F. Fitting finite mixtures of generalized linear regressions in R. Computational Statistics & Data Analysis. 2007;51(11):5247–52.
41. Guan J-H, Liu D-Y, Liu S-P. Discrete particle swarm optimization and EM hybrid approach for naive Bayes clustering. In: International Conference on Neural Information Processing, 2006;pp. 1164–1173 . Springer

Publisher's Note