Submitted to Computational Statistics and Data Analysis
arXiv: arXiv:0000.0000

VARIATIONAL ALGORITHMS FOR BICLUSTERING MODELS

By Duy Q. Vu and Murray Aitkin

University of Melbourne

Biclustering is an important tool in statistical exploratory analysis which can be used to detect latent row and column groups of different response patterns. However, few studies include covariate data directly into their biclustering models to explain these variations. In this study, we describe a model-based biclustering framework that considers both stochastic block structures and covariate effects. This introduction of covariate data together with a large number of latent variables makes the estimation task challenging. We address this problem by proposing approximation estimation algorithms derived from the variational generalized expectation-maximization (EM) framework where the goal is to increase, rather than maximize, the likelihood lower bound in both E and M steps. More specifically, we not only derive a new minorization-maximization (MM) update in the E step to model both discrete and continuous responses, but also introduce MM updates for the M step to handle high-dimensional covariate structures. The advantage of the MM principle is then demonstrated through a comparison experiment between the new MM update in the E step with the fixed-point update on two large datasets. Finally, the utility of the proposed biclustering framework is demonstrated through three block modelling applications in model-based collaborative filtering, network modelling, and microarray analysis.

1. Introduction. We consider a random matrix $\boldsymbol{Y} = \left[Y_{ij}\right] \in \Re^{n \times m}$ where observed values y_{ij} can be binary, count, or real depending on the applications involved. Our biclustering task is to simultaneously arrange rows and columns into groups of similar response patterns (Mirkin, 1996). In collaborative filtering, for example, Y_{ij} is a binary variable equal to 1 if user i liked movie j; and the modelling task is to cluster users into groups of similar reviewing patterns, while concurrently detecting movie groups with similar attractiveness levels (Su and Khoshgoftaar, 2009). Another example is from bipartite network modelling where Y_{ij} is the number of posts from user i to topic j in a Web forum; and the goal is to cluster users and topics into blocks of similar activity patterns (Borgatti and Everett, 1997; Freeman, 2003; Doreian, Batagelj and Ferligoj, 2004).

The biclustering task that we are considering is also called co-clustering (Hanisch et al., 2002), and latent or stochastic block modelling (Arabie, Boorman and Levitt, 1978). The common assumption made by these models is that each row or column belongs exclusively to only one row or column group in contrast with other overlapping biclustering models where each row or column can also belong to more than one group or none (Cheng and Church, 2000; Lazzeroni and Owen, 2002). These overlapping biclustering models and their applications in microarray analysis are discussed in detail by Madeira and Oliveira (2004). In this paper, we consider only the non-overlapping group assumption and make three significant contributions to stochastic block modelling of bipartite datasets.

Keywords and phrases: biclustering, stochastic block models, EM algorithms, generalized EM algorithms, variational EM algorithms, MM algorithms

First, we describe a biclustering framework that can jointly model latent block structures and covariate effects. Although many biclustering models have been proposed in the literature, only few take into account the covariate effects on response patterns (Madeira and Oliveira, 2004). Moreover, if considered, these covariates are only controlled for in an indirect manner. Their effects are either removed before biclustering algorithms are applied to the resulting residuals (Flynn and Perry, 2012) or only estimated after the group structures are obtained (Wang, Print and Crampin, 2013). In addition, our direct consideration of covariate structures can also help to reduce the latent class search space. The best biclustering model tends to have a smaller number of row and column groups since a large amount of variation is already explained by these observed covariates.

Our second contribution is a novel application of the variational generalized EM (VGEM) framework (Vu, Hunter and Schweinberger, 2013) to the estimation of the biclustering models. Based on the minorization-maximization principle (Hunter and Lange, 2004), our estimation algorithms seek only to increase, rather than maximize, the likelihood lower bound in both E and M steps of each variational EM iteration. More specifically, we first extend the variational generalized E step in (Vu, Hunter and Schweinberger, 2013) to deal with double latent class structures and handle both discrete and continuous responses. A comparison experiment between this new MM update and the popular fixed-point update is then carried out to demonstrate its superior performance. Furthermore, in contrast with Vu, Hunter and Schweinberger (2013) who only considered the MM principle in the variational E step, we derive new MM updates for the variational M step so that a large number of covariate effects can be estimated. These biclustering algorithms are complete examples of how the VGEM framework can be fully operationalized in practice.

Finally, the broad application of our biclustering framework and estimation algorithms is demonstrated through three different application datasets. In the *MovieLens 100K* dataset, for example, we allow the effects of 3 row and 18 column covariates to vary over columns and rows respectively, which results in the total number of 22,020 coefficients to estimate. Consequently, this application not only illustrates the important role of covariate data in model selection, but also shows how the MM principle can help to estimate such a large number of parameters. In these examples, we have also explored the application of the bootstrapping procedure in network analysis (Hunter, Goodreau and Handcock, 2008) to biclustering model diagnostics. Besides model selection criteria, these goodness-of-fit plots can provide us with a powerful visual tool for model checking.

The rest of this paper is organized as follows. In Section 2, three example datasets from different areas are discussed to motivate the application of biclustering models. The general biclustering framework is then presented in Section 3 followed by the derivation of VGEM estimation algorithms in Section 4. Section 5 shows the advantage of this VGEM framework by comparing the MM update with the FP update on two empirical datasets of both discrete and continuous responses. In Section 6, the broad application of the biclustering framework and MM algorithms is then demonstrated with three analyses of example datasets in Section 2.

2. Data. To motivate the application of the biclustering framework, we consider three real datasets corresponding to binary, count, and real response types, respectively. They also illustrate the application of the biclustering framework in three different areas: model-based collaborative filtering, network modelling, and microarray analysis.

MovieLens 100K. The dataset contains 100,000 ratings from 943 users of 1682 movies (Herlocker et al., 1999) and were retrieved from the website `http://movielens.umn.edu`. We limit our interest in this dataset to detecting viewing patterns across users and movies. Since most users rate movies only after watching them, we convert all non-zero ratings to 1 while assigning 0 to

entries without ratings. This results in a binary matrix of 943 rows and 1682 columns. The dataset also comes with 2 row covariates for age and gender, and 18 binary column covariates for movie genres. The first, second, and third quantiles of age are 25, 31, and 43, respectively; while the percentage of female reviewers is 29%. Table 1 shows the distribution of movies across different movie genres. It is important to note that movie genres are not exclusive, i.e., a movie might have multiple genres. For example, the movie *Toy Story* was simultaneously classified into three genres *Animation*, *Children*, and *Comedy*. Our modelling goals are to both extract the latent class structures of users and movies, and estimate the effects of these covariates on viewing patterns. For those readers who are interested in the modelling of rating scores, the discussion on ad-mixture models in (Zhou and Lange, 2009) provides a starting point.

Action	Adventure	Animation	Children	Comedy	Crime
15%	8%	2%	7%	30%	6%
Documentary	Drama	Fantasy	Film-Noir	Horror	Musical
3%	43%	1%	1%	5%	3%
Mystery	Romance	Sci-Fi	Thriller	War	Western
4%	15%	6%	15%	4%	2%

Table 1: The percentages of movies in each genre in the *MovieLens 100K* dataset.

UC-Irvine Forum. The original dataset, which can be downloaded from the website `http://toreopsahl.com/datasets`, was obtained from an online forum of 889 students at University of California, Irvine, in 2004 (Opsahl, 2013). Within this forum, students created 552 discussion topics; and they could post broadcast messages to any topic of which they were a member. There are 33,720 broadcast messages in total. Unfortunately, there is no information on topic memberships of these students. For this dataset, we are interested in detecting posting patterns across students and topics by converting raw data into a 899×522 matrix of posting counts between students and topics. The resulting matrix is extremely sparse, only 7,089 out of 899*552 = 496,248 entries taking non-zero values. Without considering the latent class structures, the log-likelihoods of the Poisson model and the zero-inflated Poisson (ZIP) model are -176,541 and -70,374, respectively. Because of this large improvement in the likelihood and the constraint that only active members could post to a topic, the ZIP distribution will be used to model count responses in the biclustering analysis. Covariates on users or topics are not provided with this forum dataset.

AGEMAP. This dataset is used to demonstrate our extension of the VGEM framework to continuous responses. The matrix contains the expression levels of 39 mice across 17,864 genes of two tissue types: cerebellum and cerebrum (Zahn et al., 2007) which can be downloaded from the website `http://cmgm.stanford.edu/~kimlab/aging_mouse/`. It also comes with 2 covariates on mice: age and gender. There are 4 different age groups of 1, 6, 16, and 24 months. Each combination of age and gender has 5 mice, except the group of male mice at 24 months only has 4 samples. In this application, our exploratory goal is to group samples and genes into blocks of similar expression patterns.

3. Latent class biclustering models. To model the block structure of rows and columns, we first assume that each row or column can be assigned into one of K or L latent groups, respectively. Let $\boldsymbol{Z}_i$ and $\boldsymbol{W}_j$ denote the memberships indicators of the ith row and jth column with multinomial

distributions:

$$\boldsymbol{Z}_i|\alpha_1,\ldots,\alpha_K \overset{\text{iid}}{\sim} \text{Multinomial}(1;\alpha_1,\ldots,\alpha_K),$$
$$\boldsymbol{W}_j|\beta_1,\ldots,\beta_L \overset{\text{iid}}{\sim} \text{Multinomial}(1;\beta_1,\ldots,\beta_L).$$

Conditioning on these row and column latent variables, $\boldsymbol{Z} = \left[\boldsymbol{Z}_i\right]_{i=1}^{n}$ and $\boldsymbol{W} = \left[\boldsymbol{W}_j\right]_{j=1}^{m}$, and the covariates $\boldsymbol{x}$, the Y_{ij} are assumed to be independent. This implies a factorized form of the conditional distribution of $\boldsymbol{Y}$:

$$P_{\boldsymbol{\theta}}(\boldsymbol{Y}=\boldsymbol{y}|\boldsymbol{x},\boldsymbol{Z}=\boldsymbol{z},\boldsymbol{W}=\boldsymbol{w}) \quad = \quad \prod_{i=1}^{n}\prod_{j=1}^{m} P_{\boldsymbol{\theta}}(Y_{ij}=y_{ij}|\boldsymbol{x},\boldsymbol{z},\boldsymbol{w}) \tag{3.1}$$

where $\boldsymbol{\theta}$ is the vector of model parameters. The conditional distribution of each matrix element or edge Y_{ij} in turn can be modelled by exponential families (Barndorff-Nielsen, 1978). Their distributions can be expressed in the general form

$$P_{\boldsymbol{\theta}}(Y_{ij}=y_{ij}|\boldsymbol{x},\boldsymbol{z},\boldsymbol{w}) \quad = \quad \exp[\boldsymbol{h}(\boldsymbol{\theta})^{\top}\boldsymbol{g}(\boldsymbol{x},y_{ij}) - \psi(\boldsymbol{\theta})], \tag{3.2}$$

where $\psi(\boldsymbol{\theta})$ is the normalizing constant. In the discussion below, this parameterization is illustrated for three response types including binary, count, and real-valued.

3.1. *Binary observations.* In model-based collaborative filtering, Y_{ij} can be a binary random variable to model the event that user i *reviewed* or *liked* movie j. Without covariate data on users and movies, a double latent class Rasch model (Lazarsfeld and Neil, 1968; Goodman, 1974) can be a good model that allows us to classify users into K activeness groups and movies into L attractiveness groups. In this case, the conditional probabilities are of the form:

$$P_{\boldsymbol{\theta}}(Y_{ij}=y_{ij}|z_{ik}=1,w_{jl}=1) \quad \propto \quad \exp\left[y_{ij}(\phi_k+\gamma_l)\right] \tag{3.3}$$

where the model parameter $\boldsymbol{\theta} = (\phi_1,\ldots,\phi_K,\gamma_1,\ldots,\gamma_L)$. Under this parameterization, ϕ_k models the activeness of user group k while γ_l models the attractiveness of movie group l.

From the perspective of bipartite network analysis, this model can be considered as a mixture model of bipartite exponential random graph models (ERGMs) (Wang et al., 2009) where sufficient network statistics are row and column degree sequences. Although more complex network configurations such as the number of 4-cycles (Wang, Pattison and Robins, 2013) are not considered, the latent class structures on both rows and columns allow us to introduce a weak dependence while keeping the computation feasible for large datasets (Daudin, Picard and Robin, 2008; Mariadassou, Robin and Vacher, 2010; Salter-Townshend and Murphy, 2013). Section 4.1 discusses in detail how this double latent class structure implies the dependence among Y_{ij}.

Besides the block structure, observed covariates if available should be added into the model since a large amount of response variation can be explained by observed covariates on rows, columns, or edges. From the perspective of recommender systems, this covariate extension can be considered as a hybrid model-based approach where both collaborative filtering and content-based filtering are employed (Su and Khoshgoftaar, 2009). The observed data can include row covariates $\boldsymbol{r}_i(\boldsymbol{x})$ such as age and gender of user i, or column covariates $\boldsymbol{c}_j(\boldsymbol{x})$ such as genres of movie j. Effects of row covariates can be allowed to change across columns while effects of column covariates can also be row-varying. The model also does not exclude the possiblity to model effects of edge covariates

$\boldsymbol{g}_{ij}(\boldsymbol{x})$. Examples of these edge covariates are binary indicators for the platforms where user i can watch movie j such as DVD mailing, Xbox, or online streaming. These edge effects can be fixed across rows and columns. These example covariate structures result in a more general form for the conditional probabilities as follows:

$$P_{\boldsymbol{\theta}}(Y_{ij} = y_{ij}|\boldsymbol{x}, z_{ik} = 1, w_{jl} = 1) \propto \exp\Big[y_{ij}\big(\phi_k + \gamma_l + \boldsymbol{\eta}_i{}^\top \boldsymbol{c}_j(\boldsymbol{x}) + \boldsymbol{\varphi}_j{}^\top \boldsymbol{r}_i(\boldsymbol{x}) + \boldsymbol{\nu}^\top \boldsymbol{g}_{ij}(\boldsymbol{x})\big)\Big] \tag{3.4}$$

where $\boldsymbol{\theta} = (\boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\varphi}, \boldsymbol{\nu})$. It is also possible that covariate effects are fixed across row and column latent classes, i.e. we can constrain:

$$\begin{aligned} \boldsymbol{\eta}_{i|z_{ik}=1} &= \boldsymbol{\eta}_k \\ \boldsymbol{\varphi}_{j|w_{jl}=1} &= \boldsymbol{\varphi}_l. \end{aligned} \tag{3.5}$$

For the biclustering model (3.4), if the numbers of rows and columns are large, the estimation task is very challenging. Optimization methods based on matrix inversion are not feasible due to the large number of model parameters. Section 4.3.2 will discuss how this estimation challenge can be tackled with the MM principle.

3.2. *Count observations.* In bipartite network modelling (Borgatti and Everett, 1997), Y_{ij} can be a count random variable for the number of comments that user i posted to topic j in a forum (Opsahl, 2013); or the number of times that user i checked into place j in a location-based network service (Cho, Myers and Leskovec, 2011). These network datasets tend to be very sparse, and the zero-inflated Poisson or negative binomial distribution can be used to model these observations (Greene, 1994). In this paper, we focus on the zero-inflated Poisson (ZIP) distribution which is composed of two components corresponding to different zero generating processes (Lambert, 1992). The first process is modelled by a Bernoulli distribution to generate structural zeros. Intuitively, these structural zeros can be non-exposure cases where users are not eligible to comment on topics; or users can not go to places simply because of their geographic constraints. On the other hand, the second process governed by a Poisson distribution generates counts for exposure cases. Some of the Poisson counts can be zero for those topics or places that are simply not attractive to users. In summary, we consider the conditional probabilities for zero-inflated counts of the form:

$$\begin{aligned} P_{\boldsymbol{\theta}}(Y_{ij} = y_{ij}|\boldsymbol{x}, z_{ik} = 1, w_{jl} = 1) &= \big[(1 - \rho_{kl}) + \rho_{kl} f\big(0|\lambda_{ij;kl}(\boldsymbol{\theta}, \boldsymbol{x})\big)\big]^{I(y_{ij}=0)} \\ &\quad \times \big[\rho_{kl} f\big(y_{ij}|\lambda_{ij;kl}(\boldsymbol{\theta}, \boldsymbol{x})\big)\big]^{I(y_{ij}>0)} \end{aligned} \tag{3.6}$$

where $f(y|\lambda)$ is the probability mass function of a Poisson random variable with rate λ. The rate λ in turn can be linked to observed covariates through a log-linear function:

$$\log\ \lambda_{ij;kl}(\boldsymbol{\theta}, \boldsymbol{x}) = \mu_{kl} + \boldsymbol{\eta}_i{}^\top \boldsymbol{c}_j(\boldsymbol{x}) + \boldsymbol{\varphi}_j{}^\top \boldsymbol{r}_i(\boldsymbol{x}) + \boldsymbol{\nu}^\top \boldsymbol{g}_{ij}(\boldsymbol{x}) \tag{3.7}$$

where row covariates $\boldsymbol{r}_i(\boldsymbol{x})$ can be user age or gender; column covariates $\boldsymbol{c}_j(\boldsymbol{x})$ model topic categories such as classes or football games; and edge covariates $\boldsymbol{g}_{ij}(\boldsymbol{x})$ might be binary indicators for keyword matching between user i and topic j. Under this parameterization, $e^{\mu_{kl}}$ is the expected count of the bicluster (k, l) when all observed covariates are zero.

3.3. *Real observations.* In microarray analysis, Y_{ij} can be a normal or log-normal random variable for the expression of gene j on sample i. In political network analysis, Y_{ij} can model the amount of money that organization i donated to candidate j. One possibility for latent block modelling is to assume both mean responses μ_{kl} and variances σ^2_{kl} varying across all biclusters. Covariate data can also be considered similar to the above binary and count cases. For example, if information on biological samples such as age and gender is available, they can be modelled with column-varying effects. In this case, the conditional probabilities are of the form:

$$P_{\boldsymbol{\theta}}(Y_{ij} = y_{ij}|\boldsymbol{x}, z_{ik} = 1, w_{jl} = 1) = \frac{1}{\sigma_{kl}\sqrt{2\pi}} \exp\left\{ - \frac{\left[y_{ij} - (\mu_{kl} + \boldsymbol{\varphi}_j{}^\top \boldsymbol{r}_i(\boldsymbol{x}))\right]^2}{2\sigma^2_{kl}} \right\} \tag{3.8}$$

4. Approximate maximum likelihood estimation. In this section, we first explain why the classical EM algorithm can not be used for the double latent class models defined by the conditional distribution (3.1). We then present the variational EM algorithm for the biclustering task and discuss its drawbacks. Finally, we derive our biclustering algorithms based on the variational generalized EM framework (Vu, Hunter and Schweinberger, 2013).

4.1. *Network dependence.* The classical EM algorithm (Dempster, Laird and Rubin, 1977) can not be applied to the biclustering task due to network dependence among rows and columns imposed by the block structure. Without loss of generality, we discuss this dependence with an example of binary matrix $\boldsymbol{Y}$. First, we assume that the matrix is generated from a bicluster model including two row and two column groups. Probabilities $\pi_{1,1}$ and $\pi_{2,2}$ for observing non-zero values in biclusters $(1,1)$ and $(2,2)$ are 1 and $p \in (0,1)$, respectively. However, non-zero response probabilities in two off-diagonal biclusters, $(1,2)$ and $(2,1)$, are 0:

$$\left[\begin{array}{c|cc} \pi & 1 & 2 \\ \hline 1 & 1.0 & 0.0 \\ 2 & 0.0 & p \end{array}\right]$$

If we observe $Y_{ih} = Y_{ig} = 1$, there are only two possible cluster assigments for the row i and the columns h, g. They can be members of either bicluster $(1,1)$ or $(2,2)$ only. Their group assignment, however, does not depend only on observed edges from the row i. For example, if we observe $Y_{jh} = 1$ and $Y_{jg} = 0$ for another row $j \neq i$, the row i and the columns h, g must belong to the bicluster (2,2). The reason is that if the columns h and g were assigned into the column cluster 1, Y_{jh} and Y_{jg} should both receive the same values 0 or 1 which contradicts our observations.

The above example illustrates how the biclustering framework helps to replace the independence assumption in modelling bipartite data by a more realistic conditional independence one. Although this alternative is less strong than the Markov dependence of full bipartite exponential random graph models (Wang et al., 2009; Wang, Pattison and Robins, 2013), it allows us to derive estimation algorithms that can be scaled to large datasets. The example also explains why the classical EM algorithm can not be used in the biclustering task. Rather than being independent of all other rows or columns, the conditional membership expectation of a row or column is dependent on the whole network. As an alternative, we consider using the variational inference approach (Wainwright and Jordan, 2008) in the next section.

4.2. *Variational EM algorithm.* Let $A(\boldsymbol{z}, \boldsymbol{w}) \equiv P(\boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{W} = \boldsymbol{w})$ be an auxiliary distribution on the same support as the latent variables $\boldsymbol{Z}$ and $\boldsymbol{W}$. Using Jensen's inequality, we can bound the

log likelihood as follows (Wainwright and Jordan, 2008):

$$\begin{aligned}\log P_{\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta}}(\boldsymbol{Y}=\boldsymbol{y}|\boldsymbol{x}) &= \log \sum_{\boldsymbol{z},\boldsymbol{w}} \frac{P_{\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta}}(\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{Z}=\boldsymbol{z},\boldsymbol{W}=\boldsymbol{w}|\boldsymbol{x})}{A(\boldsymbol{z},\boldsymbol{w})} A(\boldsymbol{z},\boldsymbol{w}) \\ &\geq \sum_{\boldsymbol{z},\boldsymbol{w}} \left[\log \frac{P_{\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta}}(\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{Z}=\boldsymbol{z},\boldsymbol{W}=\boldsymbol{w}|\boldsymbol{x})}{A(\boldsymbol{z},\boldsymbol{w})}\right] A(\boldsymbol{z},\boldsymbol{w}) \\ &= E_A[\log P_{\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta}}(\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{Z}=\boldsymbol{z},\boldsymbol{W}=\boldsymbol{w}|\boldsymbol{x})] - E_A[\log A(\boldsymbol{Z},\boldsymbol{W})],\end{aligned} \tag{4.1}$$

where the complete log likelihood is given by:

$$\begin{aligned}\log P_{\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta}}(\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{Z}=\boldsymbol{z},\boldsymbol{W}=\boldsymbol{w}|\boldsymbol{x}) &= \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log(\alpha_k) + \sum_{j=1}^{m}\sum_{l=1}^{L} w_{jl}\log(\beta_l) \\ &\quad + \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{K}\sum_{l=1}^{L} z_{ik}w_{jl}\log \pi_{y_{ij};kl}(\boldsymbol{\theta},\boldsymbol{x})\end{aligned} \tag{4.2}$$

where $\pi_{y_{ij};kl}(\boldsymbol{\theta},\boldsymbol{x})$ denotes the conditional probabilities (3.2).

To make inference tractable, we consider the fully factorized auxiliary variational distribution with two sets of auxiliary parameters $\tilde{\boldsymbol{z}} = (\tilde{\boldsymbol{z}}_1,\ldots,\tilde{\boldsymbol{z}}_n)$ and $\tilde{\boldsymbol{w}} = (\tilde{\boldsymbol{w}}_1,\ldots,\tilde{\boldsymbol{w}}_m)$:

$$A(\boldsymbol{z},\boldsymbol{w}) = P_{\tilde{\boldsymbol{z}},\tilde{\boldsymbol{w}}}(\boldsymbol{Z}=\boldsymbol{z},\boldsymbol{W}=\boldsymbol{w}) = \left(\prod_{i=1}^{n} P_{\tilde{\boldsymbol{z}}_i}(\boldsymbol{z}_i)\right)\left(\prod_{j=1}^{m} P_{\tilde{\boldsymbol{w}}_j}(\boldsymbol{w}_j)\right), \tag{4.3}$$

where the marginal auxiliary distributions $P_{\tilde{\boldsymbol{z}}_i}(\boldsymbol{z}_i)$ and $P_{\tilde{\boldsymbol{w}}_j}(\boldsymbol{w}_j)$ are $Multinomial(1;\tilde{z}_{i1},\ldots,\tilde{z}_{iK})$ and $Multinomial(1;\tilde{w}_{j1},\ldots,\tilde{w}_{jL})$, respectively. In this case, the variational lower bound may be written as

$$\begin{aligned}LB(\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta},\tilde{\boldsymbol{z}},\tilde{\boldsymbol{w}}|\boldsymbol{x}) &= \sum_{i=1}^{n}\sum_{k=1}^{K} \tilde{z}_{ik}\log(\alpha_k) + \sum_{j=1}^{m}\sum_{l=1}^{L} \tilde{w}_{jl}\log(\beta_l) \\ &\quad + \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{K}\sum_{l=1}^{L} \tilde{z}_{ik}\tilde{w}_{jl}\log \pi_{y_{ij};kl}(\boldsymbol{\theta},\boldsymbol{x}) \\ &\quad - \sum_{i=1}^{n}\sum_{k=1}^{K} \tilde{z}_{ik}\log \tilde{z}_{ik} - \sum_{j=1}^{m}\sum_{l=1}^{L} \tilde{w}_{jl}\log \tilde{w}_{jl}\end{aligned} \tag{4.4}$$

Approximate maximum likelihood estimates of $\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta},\tilde{\boldsymbol{z}}$, and $\tilde{\boldsymbol{w}}$ can be obtained by maximizing the lower bound in (4.4) using the variational EM algorithm (Govaert and Nadif, 2008; Shan and Banerjee, 2008; Airoldi et al., 2008). In the E step of the $t+1$ iteration, $\tilde{\boldsymbol{z}}$ and $\tilde{\boldsymbol{w}}$ can be computed by fixed-point updates while keeping other parameters fixed at values of the previous iteration:

$$\begin{aligned}\tilde{z}_{ik}^{(t+1)} &\propto \alpha_k^{(t)} \prod_{j=1}^{m}\prod_{l=1}^{L} \left[\pi_{y_{ij};kl}(\boldsymbol{\theta}^{(t)},\boldsymbol{x})\right]^{\tilde{w}_{jl}^{(t)}} \\ \tilde{w}_{jl}^{(t+1)} &\propto \beta_l^{(t)} \prod_{i=1}^{n}\prod_{k=1}^{K} \left[\pi_{y_{ij};kl}(\boldsymbol{\theta}^{(t)},\boldsymbol{x})\right]^{\tilde{z}_{ik}^{(t)}}.\end{aligned} \tag{4.5}$$

In the M step, $\boldsymbol{\theta}, \boldsymbol{\alpha}$, and $\boldsymbol{\beta}$ can be estimated by maximizing the lower bound (4.4) while keeping $\tilde{\boldsymbol{z}}$ and $\tilde{\boldsymbol{w}}$ fixed at $\tilde{\boldsymbol{z}}^{(t+1)}$ and $\tilde{\boldsymbol{w}}^{(t+1)}$, respectively.

The potential problems with fixed-point updates in the E step are their slow empirical convergence and the possibility of local maxima due to the highly non-concave nature of the log likelihood. Moreover, for the biclustering models discussed in Section 3, maximizing the lower bound at the M step is also challenging due to the large number of parameters when high-dimensional covariate structures are included. The next section proposes variational generalized EM algorithms to address these issues.

4.3. *Variational generalized EM algorithm.* Vu, Hunter and Schweinberger (2013) investigated the variational generalized EM framework where the principle in both E and M steps is to increase the lower bound rather than to maximize it. However, they focused only on multinomial responses and have not considered using the MM principle in the M step to fully explore its potential. In this section, we first adjust their MM update for the E step in so that the biclustering algorithms can handle both discrete and continuous responses. To estimate coefficients of high dimensional covariate structures, we then develop new MM updates for the M step by constructing minorizing functions of the lower bound (4.4). All inequalities used in the construction of these minorizers are listed in Appendix A.

4.3.1. *Variational generalized E-step.* Similar to (Vu, Hunter and Schweinberger, 2013), the arithmetic-geometric mean inequality (A.1) can be used to construct a minorizer for the lower bound $LB(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \tilde{\boldsymbol{z}}, \tilde{\boldsymbol{w}} | \boldsymbol{x})$ of discrete responses in the variational E step. This implies the following inequality:

$$\log \pi_{y_{ij};kl}(\boldsymbol{\theta}, \boldsymbol{x})\ \tilde{z}_{ik}\, \tilde{w}_{jl} \quad \geq \quad \log \pi_{y_{ij};kl}(\boldsymbol{\theta}, \boldsymbol{x}) \left(\tilde{z}_{ik}^2 \frac{\tilde{w}_{jl}^{(t)}}{2\, \tilde{z}_{ik}^{(t)}} + \tilde{w}_{jl}^2 \frac{\tilde{z}_{ik}^{(t)}}{2\, \tilde{w}_{jl}^{(t)}} \right) \tag{4.6}$$

with equality if $\tilde{z}_{ik} = \tilde{z}_{ik}^{(t)}$ and $\tilde{w}_{jl} = \tilde{w}_{jl}^{(t)}$. However, this inequality only holds for a discrete random variable since its mass probability function $\pi_{y_{ij};kl}(\boldsymbol{\theta}, \boldsymbol{x})$ is less than or equal to one. For continuous observations, it is possible that $\log \pi_{y_{ij};kl}(\boldsymbol{\theta}, \boldsymbol{x}) > 0$ because the density function $\pi_{y_{ij};kl}(\boldsymbol{\theta}, \boldsymbol{x})$ can be greater than one. In this case, the inequality (A.2) can be used which implies:

$$\begin{aligned} \log \pi_{y_{ij};kl}(\boldsymbol{\theta}, \boldsymbol{x})\ \tilde{z}_{ik}\, \tilde{w}_{jl} \quad \geq \quad & \log \pi_{y_{ij};kl}(\boldsymbol{\theta}, \boldsymbol{x}) \big[-\frac{1}{2}(\tilde{z}_{ik}^2 + \tilde{w}_{jl}^2) - \frac{1}{2}(\tilde{z}_{ik}^{(t)} + \tilde{w}_{jl}^{(t)})^2 \\ & + (\tilde{z}_{ik}^{(t)} + \tilde{w}_{jl}^{(t)})(\tilde{z}_{ik} + \tilde{w}_{jl}) \big] \end{aligned} \tag{4.7}$$

with equality if $\tilde{z}_{ik} = \tilde{z}_{ik}^{(t)}$ and $\tilde{w}_{jl} = \tilde{w}_{jl}^{(t)}$.

In addition, using the log concavity inequality (A.3) for both functions $\log \tilde{z}_{ik}$ and $\log \tilde{w}_{jl}$, we can obtain the general minorizer of the variational lower bound in the E step for both response types:

$$\begin{aligned} LB(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{w}} | \boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \quad & \geq \quad Q(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{w}} | \boldsymbol{x}, \boldsymbol{\theta}^{(t)}, \boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t)}, \tilde{\boldsymbol{z}}^{(t)}, \tilde{\boldsymbol{w}}^{(t)}) \\ & = \quad \sum_{i=1}^{n} \sum_{k=1}^{K} \left(\tilde{z}_{ik}^2\ A^{(t)}(i,k) + \tilde{z}_{ik} S^{(t)}(i,k) \right) \\ & + \quad \sum_{j=1}^{m} \sum_{l=1}^{L} \left(\tilde{w}_{jl}^2\ B^{(t)}(j,l) + \tilde{w}_{jl} T^{(t)}(j,l) \right) \end{aligned} \tag{4.8}$$

subject to probability constraints on $\tilde{\boldsymbol{z}}$ and $\tilde{\boldsymbol{w}}$. All coefficients of these quadratic minorizing functions are computed based on parameter estimates of the previous iteration; and their derivations are discussed in detail in Appendix B.1. Estimates $\tilde{\boldsymbol{z}}^{(t+1)}$ and $\tilde{\boldsymbol{w}}^{(t+1)}$, therefore, can be obtained by solving $n+m$ quadratic programming problems separately for each latent row variable $\tilde{\boldsymbol{z}}_i$ or column variable $\tilde{\boldsymbol{w}}_j$ (Stefanov, 2004). These MM updates in the E step can be summarized as follows:

Variational generalized E-step for $\tilde{\boldsymbol{z}}_i$: solve the K-dimensional quadratic problem

$$\begin{aligned} &\sum_{k=1}^{K}\left(\tilde{z}_{ik}^2\, A^{(t)}(i,k)+\tilde{z}_{ik}S^{(t)}(i,k)\right) \\ \text{subject to}\quad &\tilde{z}_{ik}\geq 0\ \forall k\ \text{ and }\ \sum_{k=1}^{K}\tilde{z}_{ik}=1; \end{aligned} \tag{4.9}$$

Variational generalized E-step for $\tilde{\boldsymbol{w}}_j$: solve the L-dimensional quadratic problem:

$$\begin{aligned} &\sum_{l=1}^{L}\left(\tilde{w}_{jl}^2\, B^{(t)}(j,l)+\tilde{w}_{jl}T^{(t)}(j,l)\right) \\ \text{subjects to}\quad &\tilde{w}_{jl}\geq 0\ \forall l\ \text{ and }\ \sum_{l=1}^{L}\tilde{z}_{jl}=1. \end{aligned} \tag{4.10}$$

Iterative Schemes. There are two schemes to carry out variational EM iterations regardless of the update methods in the E step (Govaert and Nadif, 2008). In the simultaneous row and column update scheme (SUS), both row and column latent variables $\tilde{\boldsymbol{z}}$ and $\tilde{\boldsymbol{w}}$ are updated in the E step first before the estimation of $\boldsymbol{\theta}, \boldsymbol{\alpha}$, and $\boldsymbol{\beta}$ in the M step. In the block update scheme (BUS), on the other hand, only the set of row or column latent variables is updated in each EM iteration. In the next EM iteration, the alternative set is estimated instead. Under the variational EM framework, the block update scheme is reported to have outperformed the simultaneous row and column update scheme (Govaert and Nadif, 2008). Since this result may not hold under the variational generalized EM framework, we will compare their relative performance on two binary and real-valued datasets in Section 5.

4.3.2. *Variational generalized M-step.* The estimation of $\boldsymbol{\alpha}, \boldsymbol{\beta}$, and $\boldsymbol{\theta}$ can be accomplished separately in the variational M step. First, by setting the derivatives of the lower bound (4.4) with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ equal to zeros, we can obtain the closed-form updates for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$:

$$\begin{aligned} \alpha_k^{(t+1)} &= \frac{1}{n}\sum_{i=1}^{n}\tilde{z}_{ik}^{(t+1)},\ k=1,\ldots,K; \\ \beta_l^{(t+1)} &= \frac{1}{m}\sum_{j=1}^{m}\tilde{w}_{jl}^{(t+1)},\ l=1,\ldots,L. \end{aligned} \tag{4.11}$$

Second, model parameters $\boldsymbol{\theta}$ can be estimated by maximizing the following part of the variational lower bound that contains $\boldsymbol{\theta}$:

$$LB(\boldsymbol{\theta}|\tilde{\boldsymbol{z}}^{(t+1)},\tilde{\boldsymbol{w}}^{(t+1)}) \propto \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{K}\sum_{l=1}^{L}\tilde{z}_{ik}^{(t+1)}\tilde{w}_{jl}^{(t+1)}\log\pi_{y_{ij};kl}(\boldsymbol{\theta},\boldsymbol{x}) \tag{4.12}$$

We consider MM updates in the M step for three reasons. First, the popular gradient ascent and Newton's methods are not suitable for the maximization of the objective function (4.13) due to the large number of parameters. The matrix inversion of Newton's methods is expensive while gradient ascent methods may be unstable and nontrivial to implement (Lange, 2010). On the other hand, our new MM updates do not require matrix inversion since all parameters are separated. They are also robust and simple to implement. Second, as demonstrated by previous empirical studies (Neal and Hinton, 1993; Lange, 2010), seeking only to increase the objective function rather than maximizing it in the EM algorithm can lead to faster convergence. The VGEM framework provides us a principle approach to derive such incremental updates in the M step. Finally, the separation of the resulting parameter updates also allows us to use parallel programming or GPU computing for computation acceleration. Though we do not pursue this parallelization idea further in this paper, for those interested readers, a recent investigation on MM algorithms and GPU computing can be found in (Zhou, Lange and Suchard, 2010).

In this section, we will consider only MM updates in the M step for the binary biclustering model (3.4). The derivation of MM updates for count and real responses are deferred to Appendix B. Without loss of generality, we focus on the case when the model contains only column-varying effects for row covariates. In this case, the corresponding lower bound (4.12) is of the form:

$$LB(\boldsymbol{\theta}|\tilde{\boldsymbol{z}}^{(t+1)},\tilde{\boldsymbol{w}}^{(t+1)}) \propto \sum_{i,j,k,l} \tilde{z}_{ik}^{(t+1)}\tilde{w}_{jl}^{(t+1)}\Big[y_{ij}\big(\phi_k+\gamma_l+\boldsymbol{\varphi}_j{}^{\top}\boldsymbol{r}_i(\boldsymbol{x})\big)-\log\Psi_{ij;kl}(\boldsymbol{x})\Big] \tag{4.13}$$

where $\Psi_{ij;kl}(\boldsymbol{x}) = 1+\exp\big(\phi_k+\gamma_l+\boldsymbol{\varphi}_j{}^{\top}\boldsymbol{r}_i(\boldsymbol{x})\big)$ and the sums are over all possible values of i, j, k, and l.

The concave nature of the logarithm (A.3) and the exponential convexity inequality (A.4) allow us to separate all parameters and construct the final minorizer of the form:

$$\begin{aligned} LB(\boldsymbol{\theta}|\tilde{\boldsymbol{z}}^{(t+1)},\tilde{\boldsymbol{w}}^{(t+1)}) &\geq Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)},\tilde{\boldsymbol{z}}^{(t+1)},\tilde{\boldsymbol{w}}^{(t+1)}) \\ &\propto \sum_{k=1}^{K} Q_{\boldsymbol{\phi}}(\phi_k)+\sum_{l=1}^{L} Q_{\boldsymbol{\gamma}}(\gamma_l)+\sum_{j=1}^{m}\sum_{q=1}^{p} Q_{\boldsymbol{\varphi}}(\varphi_{jq}). \end{aligned} \tag{4.14}$$

Details of this derivation and the formulas of individual minorizers $Q_{\boldsymbol{\phi}}(\phi_k)$, $Q_{\boldsymbol{\gamma}}(\gamma_l)$, and $Q_{\boldsymbol{\varphi}}(\varphi_{jq})$ are given in Appendix B.2. These individual minorizers are conditioned on the current estimates of $\boldsymbol{\theta}^{(t)}, \tilde{\boldsymbol{z}}^{(t+1)}$, and $\tilde{\boldsymbol{w}}^{(t+1)}$, which are dropped out to simplify the notations.

Since all minorizers are univariate, one-step Newton's method can be used to estimate the corresponding parameters separately (Lange, 2010, Chapter 14). More specifically, update formulas for $\phi_k, \gamma_l, \varphi_{jq}$ are of the form:

$$\begin{aligned} \phi_k^{(t+1)} &= \phi_k^{(t)} - \xi_{\phi_k} \times \bigtriangledown^2 Q_{\boldsymbol{\phi}}(\phi_k^{(t)})^{-1} \times \bigtriangledown Q_{\boldsymbol{\phi}}(\phi_k^{(t)}) \quad \text{for } k=1,\ldots,K; \\ \gamma_l^{(t+1)} &= \gamma_l^{(t)} - \xi_{\gamma_l} \times \bigtriangledown^2 Q_{\boldsymbol{\gamma}}(\gamma_l^{(t)})^{-1} \times \bigtriangledown Q_{\boldsymbol{\gamma}}(\gamma_l^{(t)}) \quad \text{for } l=1,\ldots,L; \\ \varphi_{jq}^{(t+1)} &= \varphi_{jq}^{(t)} - \xi_{\varphi_{jq}} \times \bigtriangledown^2 Q_{\boldsymbol{\varphi}}(\varphi_{jq}^{(t)})^{-1} \times \bigtriangledown Q_{\boldsymbol{\varphi}}(\varphi_{jq}^{(t)}) \quad \text{for } j=1,\ldots,m; q=1,\ldots,p. \end{aligned} \tag{4.15}$$

The step lengths $\xi_{\phi_k}, \xi_{\gamma_l}, \xi_{\varphi_{jq}}$ can be backtracked from 1.0 until the corresponding component minorizers increase if necessary.

5. Comparison experiment. To demonstrate the advantage of the MM principle for large datasets, we compare the MM and FP updates in the variational E step on two large datasets.

The first dataset is *MovieLens 1M* from http://movielens.umn.edu which contains one million reviews of 6,040 users on 3,952 movies. It was converted into a binary matrix by the same procedure applied to *MovieLens 100K* in Section 2. The second dataset is a gene expression matrix of 255 *Saccharomyces Cerevisiae* samples across 6,189 genes (Beer and Tavazoie, 2004). Comparing to the experiment in (Vu, Hunter and Schweinberger, 2013) which focused only on network datasets with multinomial responses, our experiment considers both discrete and continuous responses.

5.1. *Experiment settings.* Two different cluster settings are used to illustrate the robustness of MM updates when the number of clusters is large. Table 2 shows these cluster settings and experimental wall-times for each dataset. Although the variational updates in the E step are different, both algorithms use the same closed-form updates (B.4) and (B.9) in the M step since covariate effects are not considered.

Since the variational EM algorithms could become stuck in distinct local maxima, we use multiple different starting values to search for parameter estimates that achieve the best lower bound. We assign the values of $\tilde{\boldsymbol{z}}^{(0)}, \tilde{\boldsymbol{w}}^{(0)}$ as follows: the initial $\tilde{z}_{ik}^{(0)}$ are drawn independently randomly from a uniform distribution on $(0, 1)$, then each $\tilde{\boldsymbol{z}}_i^{(0)}$ is normalized by a constant chosen so that the elements of $\tilde{\boldsymbol{z}}_i^{(0)}$ sum to one for every i. The same initialization procedure is applied to column variational variables $\tilde{\boldsymbol{w}}^{(0)}$. After the initialization of row and column variational variables, the starting values of $\boldsymbol{\alpha}^{(0)}, \boldsymbol{\beta}^{(0)}$, and $\boldsymbol{\theta}^{(0)}$ are estimated with an M step. The algorithm then continues with the first E step, and so on. For each dataset, each cluster setting, and each algorithm, we carried out 100 random initial runs.

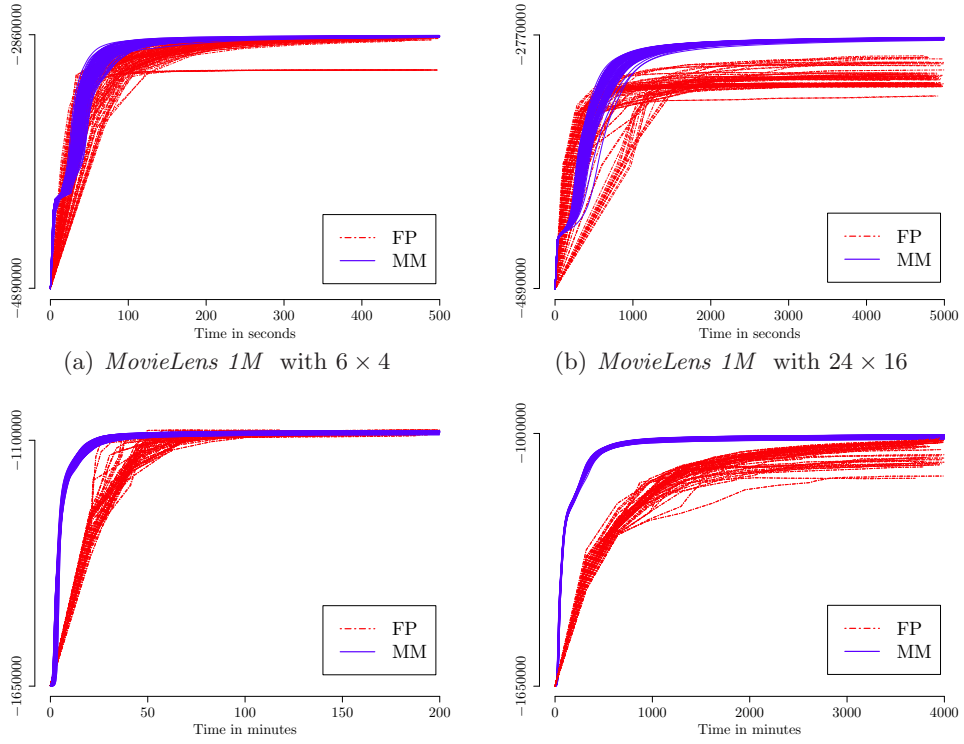
We employ three stopping criteria. The first stopping rule is to limit the number of EM iterations to 5,000. The second stopping rule is a convergence criterion where we stop the algorithm as soon as the relative change in the objective function is smaller than 10^{-10}. The last stopping rule is to set a wall-time of 24 or 96 hours on the running time depending on the cluster settings. Most of these runs were finished before reaching these wall-times.

Dataset	Size	Small Cluster Settting	Large Cluster Setting
MovieLens 1M	$6,040 \times 3,952$	6×4	24×16
		(24 hours)	(96 hours)
Saccharomyces Cerevisiae	$255 \times 6,189$	3×6	12×24
		(24 hours)	(96 hours)

Table 2: Two cluster settings are used in the comparison experiment. Matrix and cluster sizes are written as the product of the numbers of rows and columns. The wall-times are shown in parentheses.

5.2. *Experiment results.* Figure 1 shows trace plots of the lower bounds $LB(\boldsymbol{\gamma}^{(t)}, \boldsymbol{\theta}^{(t)}; \boldsymbol{\alpha}^{(t)})$ of the log-likelihood functions. Red and blue lines refer respectively to the lower bounds of the VGEM algorithms with the FP and MM updates in the E step. On both datasets, the MM method is far superior to the FP method, especially on the setting with a larger number of clusters. MM updates converge faster and are not prone to local maxima. We believe that the overall better performance of MM updates in the E step stems from the fact that the MM principle allows us to separate the variational parameters $\tilde{z}_{ik}$ and $\tilde{w}_{jl}$, which in turn results in the weak dependence of the resulting updates. In brief, our experiment results appear to agree with Vu, Hunter and Schweinberger (2013) and support the use of MM updates with a small number of random initial runs and longer wall-times that allow these runs to converge.

In this comparison experiment, we also extend the previous study (Vu, Hunter and Schweinberger, 2013) by comparing the performance of the simultaneous row and column update scheme and the block update scheme. In the block update scheme, as discussed in Section 4.3.1, only the set of row or column variational variables is updated in the E step before the estimation of model parameters in the M step. Both discrete *MovieLens 1M* and continuous *Saccharomyces Cerevisiae* datasets are used in this experiment. In Figure 2, red lines refer to the lower bound of the VGEM algorithm with the block update scheme while blue lines refer to the lower bound of the VGEM algorithm with the simultaneous row and column update scheme. Regarding local maxima, our experiment demonstrates that there is no difference between two methods. Moreover, the simultaneous row and column update scheme seems to converge faster on the continuous dataset. Overall, these results contrast with the conclusions in (Govaert and Nadif, 2008) where the block update scheme is better. We suspect that the separation of parameters and the resulting weak dependence among corresponding updates make the incremental advantage of the block update scheme unnecessary.

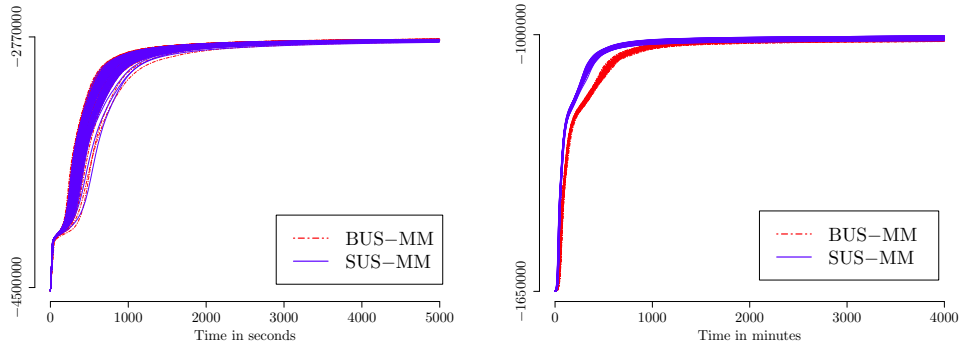


(a) *MovieLens 1M* with 6×4 (b) *MovieLens 1M* with 24×16

(c) *Saccharomyces Cerevisiae* with 3×6 (d) *Saccharomyces Cerevisiae* with 12×24

FIG 1. *The comparison results between VGEM algorithms with FP updates (4.5) and MM updates (4.9 and 4.10) in the E step on the MovieLens 1M and Saccharomyces Cerevisiae datasets.*

6. Applications. In this section, we apply the proposed biclustering models and algorithms to three example datasets in Section 2. In all analyses, to search for the best models, 10 rather than 100 random initial runs are used for each cluster setting. The run that achieves the best lower bound is then compared with the best runs of other cluster settings. Due to the large numbers of cluster settings, this small number of random initial runs helps to reduce the total computational



(a) *MovieLens 1M* with 24×16 (b) *Saccharomyces Cerevisiae* with 12×24

FIG 2. *The comparison results on the MovieLens 1M and Saccharomyces Cerevisiae datasets between two VGEM algorithms with different iterative schemes: the block update scheme (BUS-MM) and the simultaneous row and column update scheme (SUS-MM).*

load. This choice is also supported by the results in Section 5 that MM updates are robust to starting values. Moreover, the MM updates (4.9) and (4.10) for the variational E step with the simultaneous row and column update scheme are used in all analyses. We use three stopping rules similar to Section 5. Data sizes, cluster settings, and wall-times of these application analyses are shown in Table 3. Moreover, the Bayesian information criterion (BIC) is used to select the best models.

Dataset	Size	Row Clusters	Column Clusters	Wall-time
MovieLens 100K	$943 \times 1,682$	$1, \ldots, 3$	$1, \ldots, 5$	192 hours
UC-Irvine Forum	889×552	$1, \ldots, 10$	$1, \ldots, 10$	96 hours
AGEMAP	$39 \times 17,864$	3	$1, \ldots, 30$	96 hours

Table 3: The cluster settings and wall-times used in three application analyses.

6.1. *MovieLens 100K dataset with row and column covariates.* In this analysis, we use the binary response model (3.4) where row covariate effects are allowed to change over columns and column covariate effects can be varied across rows. Since there are 943 rows and 1,682 columns, 18 column covariates results in 18*943 = 16,974 row parameters while two row covariates and their interaction add 3*1,682 = 5,046 column parameters to the model. Regarding the estimation, we use one-step Newton-Raphson with the default step length 1.0 in the variational M step as discussed in Section 4.3.2.

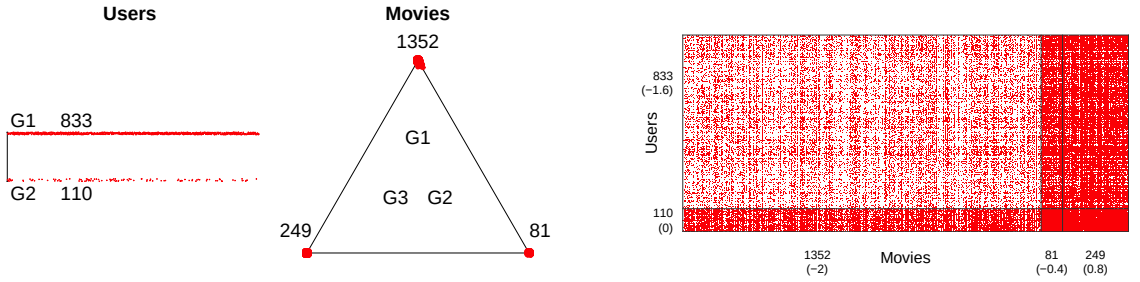
To select the best model, we vary the number of row clusters from 1 to 3 and the number of column clusters from 1 to 5. Table 4 shows BICs of all cluster settings. In a pilot study where we did not consider covariate data, BIC improved from the model with 3 row and 6 column clusters to the model with 6 row and 12 column clusters, while the best model in Table 4 has only two row groups and three column groups. The selection of smaller numbers of row and column groups can be explained by the fact that a large amount of response variation is already explained by the covariates. The remaining variation can then be captured effectively by a lower dimensional block structure. We explore futher the fitness of this model of two user groups and three movie groups in the next discussion.

Under the variational inference approach, the row and column membership posteriors of $\boldsymbol{z}_i$ and $\boldsymbol{w}_j$ can be approximated by $P_{\tilde{\boldsymbol{z}}_i}(\boldsymbol{z}_i)$ and $P_{\tilde{\boldsymbol{w}}_j}(\boldsymbol{w}_j)$, respectively. Based on these approximate

		The number of column groups				
		1	2	3	4	5
The number of row groups	1	768,594	762,546	762,340	762,004	762,161
	2	765,568	759,445	**759,081**	759,147	759,198
	3	765,692	759,590	759,372	760209	759,806

Table 4: BICs of the biclustering models on the *MovieLens 100K* dataset.

posteriors, Figure 3(a) shows the group assigments of both users and movies. All rows and columns are clearly assigned into two user groups and three movie groups, respectively. Figure 3(b) shows the rating matrix where rows and columns are re-arranged by their group assigments according to the selected model. Users are divided into two groups. The first large group of 833 users is less active in reviewing than the second smaller group of 110 users. Movies are separated into three groups of sizes 1352, 81, and 249, respectively. Movie groups further to the right are more attractive since they receive more reviews.

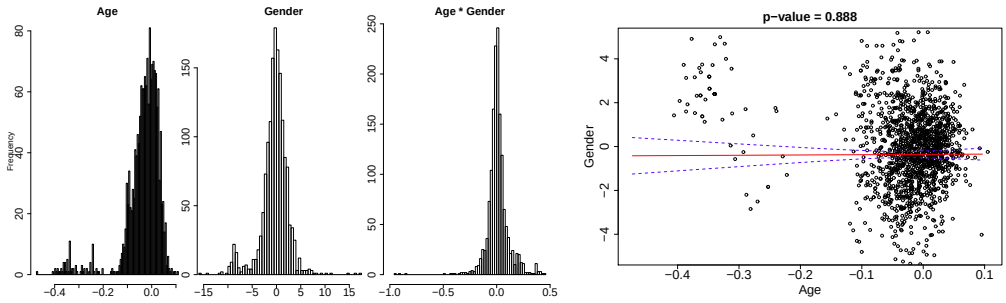


(a) Membership Posteriors (b) Biclustered Matrix

FIG 3. *(a) Membership plots for users and movies where all of them are clearly assigned into one row or column group. Red points, users and movies, are drawn closer to the group labels with the highest posterior probabilities. (b) The rating matrix MovieLens 100K is re-arranged using group assigments from the best biclustering model. Row groups 1 and 2 have 833 and 110 users, respectively. Column groups 1, 2, and 3 have 1,352, 81, and 249 movies, respectively. Values in parentheses are the estimates of activeness effects of user groups ϕ_k and attractivenss effects of movie groups γ_l.*

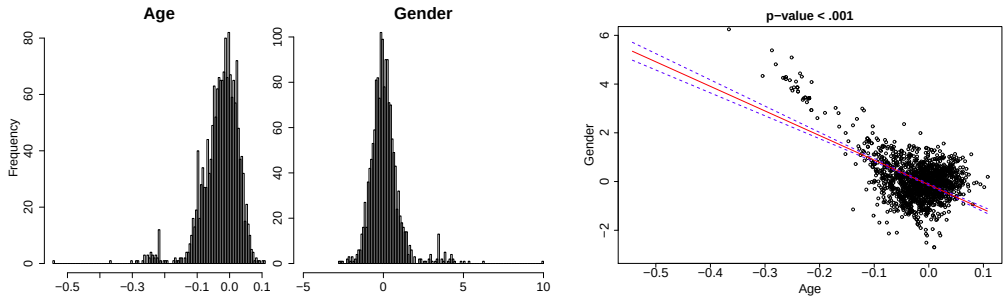
Figure 4(a) shows estimated column-varying coefficients for user age and gender covariates and their interaction which vary largely across columns. Figure 1 in Appendix C shows row-varying coefficient estimates for 18 binary column covariates on movie genres which also change considerably across rows. These large variances of estimated coefficients for row and column covariates, therefore, support our assumption of row and column varying effects. Moreover, the consideration of the age and gender interaction is neccessary since the variance of its column-varying effects is large. It also helps to remove the correlation between age and gender effects as indicated in Figure 4(b). In a preliminary analysis, we also considered biclustering models without the interaction term. The best model was also the clustering configuration with 2 row and 3 column clusters. Figure 5(a) shows the estimated coefficients of age and gender covariates while Figure 5(b) reveals a negative correlation between age and gender effects.

Finally, we explore some limitations of the proposed biclustering model from a network modelling perspective. We suspect that the model may not be able to capture higher-order network configurations such as two-paths and four-cycles (Wang et al., 2009). Figure 6 illustrates two configurations of 2-paths, User-Movie-User (UMU) and Movie-User-Movie (MUM), and the 4-cycle configuration.



(a) Column-varying coefficients (b) Coefficient correlation

FIG 4. *(a) Column-varying coefficients of age and gender covariates and their interaction term. Since 148 movies were not rated by any woman, their large estimated coefficients of the gender and interaction effects are removed from these plots. (b) The correlation plot between age and gender coefficient estimates.*



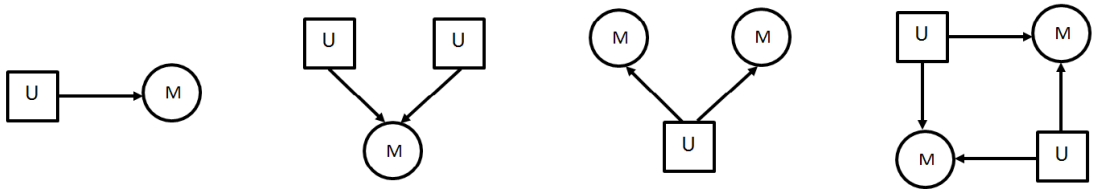
(a) Column-varying coefficients (b) Coefficient correlation

FIG 5. *(a) Column-varying coefficients of two row covariates of user age and gender. (b) The correlation plot between age and gender coefficient estimates.*

They are the important descriptive statistics that are often used to measure network dependence in bipartite data. We also consider other low-order network statistics including density, i.e. the number of edges, user and movie degree distributions. The degree of a user is the number of reviews that she or he has made. The degree of a movie is the number of ratings that it has received.

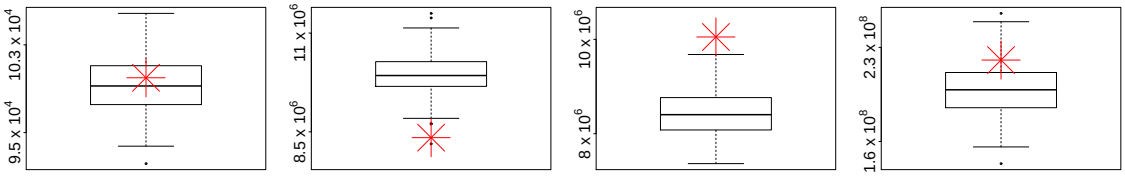
We follow the goodness of fit bootstrapping procedure for exponential random graph models (Hunter, Goodreau and Handcock, 2008; Wang et al., 2009) by generating random networks from the biclustering model using the estimated parameters. Since only users with at least 20 reviews were kept in the original dataset, we reject a network sample if the degree of any user is smaller than 20. Observed network statistics including density, the numbers of two 2-path configurations, the number of 4-cycles, and two degree distributions (shown by red asterisks) are then compared with bootstrapped values and distributions of 200 random networks (shown by boxplots) in Figures 7 and 8.

The goodness of fit of the biclustering model is good with respect to the density and 4-cycles as shown in Figures 7(a) and 7(d). In Figures 7(b) and 7(c), however, the biclustering model overestimates the number of UMU 2-paths while underestimating the number of MUM two-paths. Figure 8 shows the model goodness of fit regarding the degree distributions. The movie degree distribution is well reproduced by the model, but the lowest user degrees are underestimated. We suspect that the lack of fit of the biclustering model could be due to the truncation of the original



(a) Density (b) UMU 2-Paths (c) MUM 2-Paths (d) 4-Cycles

FIG 6. *Network configurations used to evaluate the goodness of fit of the best biclustering model. Squares and circles represent users and movies, respectively. Directed edges indicate review events from users to movies.*



(a) Density (b) UMU 2-Paths (c) MUM 2-Paths (d) 4-Cycles

FIG 7. *Goodness of fit of the best biclustering model on MovieLens 100K regarding density, two-paths statistics, and 4-cycles. Red asterisks are observed values of these network statistics while boxplots are drawn to show the distributions of their sampled values based on 200 boostrapped networks. Only sampled networks with the smallest user degrees greater than or equal to 20 were kept since this is the main property in the original dataset.*

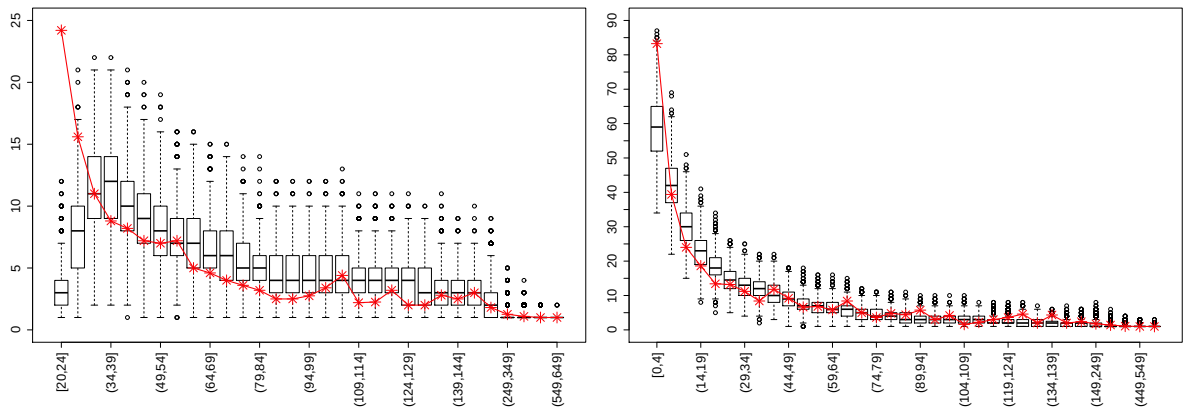
data where only users with at least 20 ratings were retained.

6.2. *UC-Irvine Forum dataset without covariates.* For the *UC-Irvine Forum* dataset with count responses, the ZIP biclustering model in Section 3.2 is used. To search for the best model, we vary the numbers of row and column clusters from 1 to 10. Regarding the estimation method, we use the closed-form MM update (B.6) in the variational M step since no covariate is considered. Table 5 shows BICs of the models with the numbers of row and column clusters ranging from 7 to 10. The best model has 9 user groups and 9 topic groups. In the discussion below, we explore this selected model in detail.

		The number of column groups			
		7	8	9	10
The number of row groups	7	105,735	105,495	105,440	105,345
	8	105,323	105,361	105,129	105,274
	9	105,410	105,223	**104,976**	105,060
	10	105,253	105,131	105,065	105,062

Table 5: BICs of different biclustering models on the *UC-Irvine Forum*dataset. Higher BIC values of smaller cluster settings are not shown.

Figure 9(a) shows the certainty of the selected model in assigning rows and columns into clusters. There are only 84.3% rows which are assigned to a group with .9 probability while 88.4% columns are classified into a group with .9 probability. This divergence in membership posteriors indicates an uncertainty in the bicluster structure. We study the model fitness further by simulating 200 count matrices from the estimated model. Figure 9(b) compares the observed count distribution (shown



(a) User Degree Distribution (b) Movie Degree Distribution

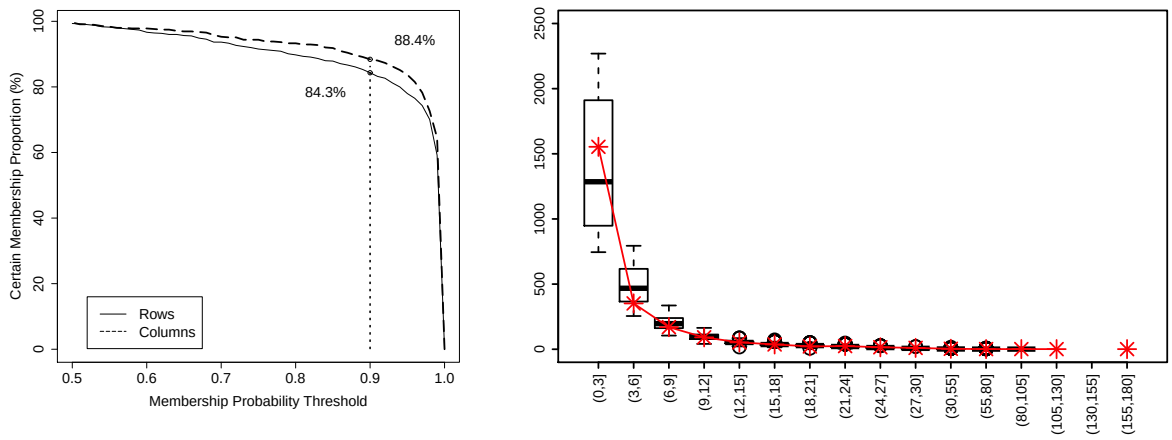
FIG 8. *Goodness of fit of the best biclustering model on MovieLens 100K regarding user and movie degree sequences. Red asterisks are observed user and movie degree distributions while boxplots represent the boostrapped distributions. Due to the long-tail nature of these distributions, degrees are combined into groups of unequal sizes. A group size of 5 is used for degrees smaller than or equal to 150. For degrees greater than 150, the group size is 100.*

by red asterisks) with the bootstrapped ones (shown by boxplots). Overall, the selected model can successfully reproduce the observed count distribution except for some extreme values greater than 100. Morever, our zero-inflated model assumption is clearly supported by the goodness of fit check since the observed zero count is well covered by the corresponding bootstrapped values in Figure 10(a). However, the selected model can not fully capture other high-order network statistics. Although two 2-path configurations, User-Topic-User (UTU) and Topic-User-Topic (TUT), can be reproduced closely by the bootstrapped samples, the 4-cycle configuration is far from being fitted.

Figure 11 shows the model parameters grouped by row and column clusters. These row and column parameter groups are sorted by the medians of their ad-mixture parameters as indicated in the ad-mixture parameter boxplots. These plots reveal an interesting nonlinear relationship between ad-mixture and rate parameters. For example, the exposure probabilities (i.e. ad-mixture parameters) of the user group 7 is much smaller then those of the user group 9; however, conditioning on exposure events, users in group 7 are more active in posting than users in group 9.

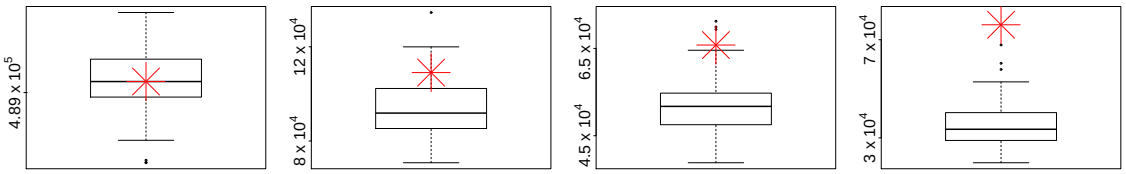
6.3. *AGEMAP dataset with row covariates.* For the *AGEMAP* dataset, we use the real-valued model (3.8) where age and gender effects are allowed to change over column genes. We also include one interaction between age and gender, which results in a total number of 3*17,864 = 53,592 column parameters. Regarding the estimation, we use the closed-form updates (B.7) and (B.8) in the variational M step discussed in Appendix B.

To select the best model, we fix the number of row clusters to 3 following the results from (Perry and Owen, 2010) and vary column clusters from 1 to 30. Figure 12(a) shows BICs of all considered cluster configurations. The lowest BIC is achieved by the model with 22 column groups, though the improvement in BIC is small when the number of column groups reaches 15. Figure 12(b) showing three row curves of mean parameters μ_{kl} also explains for this observation. For example, bicluster means of column groups 9, 10, and 11 are only slightly different, and potentially these columns could be modelled by one group. The re-arranged gene expression matrix in Figure 13 reveals no clear difference in expression levels across column groups 9, 10, and 11. A similar reasoning could also be applied to column clusters 21 and 22. Despite this uncertainty of the estimated block



(a) Membership Certainty (b) Goodness of Fit

FIG 9. *(a) The membership certainty curves for the UC-Irvine Forum dataset show the proportions of rows and columns which are assigned to groups with membership probablities greater than varying thresholds. The larger the threshold is, the smaller proportions of rows and columns are certainly classified. (b) Goodness of fit plot of the observed and bootstrapped count distributions. The bootstrapped distributions are computed based on 200 matrix samples.*



(a) Zero Counts (b) UTU 2-Paths (c) TUT 2-Paths (d) 4-Cycles

FIG 10. *Goodness of fit of the best biclustering model on UC-Irvine Forum regarding zero counts, two-paths statistics, and 4-cycles. Red asterisks are observed values of these network statistics while boxplots are drawn to show the distributions of their sampled values based on 200 boostrapped networks. In this computation of sampled network configurations, an edge between user i and topic j exists if there is at least one post between them*

structure, Figure 13 successfully demonstrates the ability of our biclustering model in detecting blocks of genes that have similar expression patterns.

Figure 14(a) are histograms of estimated column-varying coefficients for mice age and gender covariates, and their interaction from the best model with 22 column clusters. These coefficients vary largely supporting our assumption of column-varying effects. In other words, age and gender are associated with the expression levels; however, their effects vary over genes. Compared to the *MovieLens 100K*, the correlation between age and gender effects in Figure 14(b) is not fully eliminated even after the interaction term is included. Their positive correlation implies a similar effect direction of both covariates on the expression levels.

7. Discussion. Our study makes several contributions to the modelling and estimation of the biclustering problem. First, we have proposed a biclustering framework that can jointly model block structures and covariate effects. This approach is in contrast with other methods where covariate effects are adjusted only before or after the performance of the biclustering task (Flynn and Perry, 2012; Wang, Print and Crampin, 2013). Different possibilities of parameterization for covariate ef-

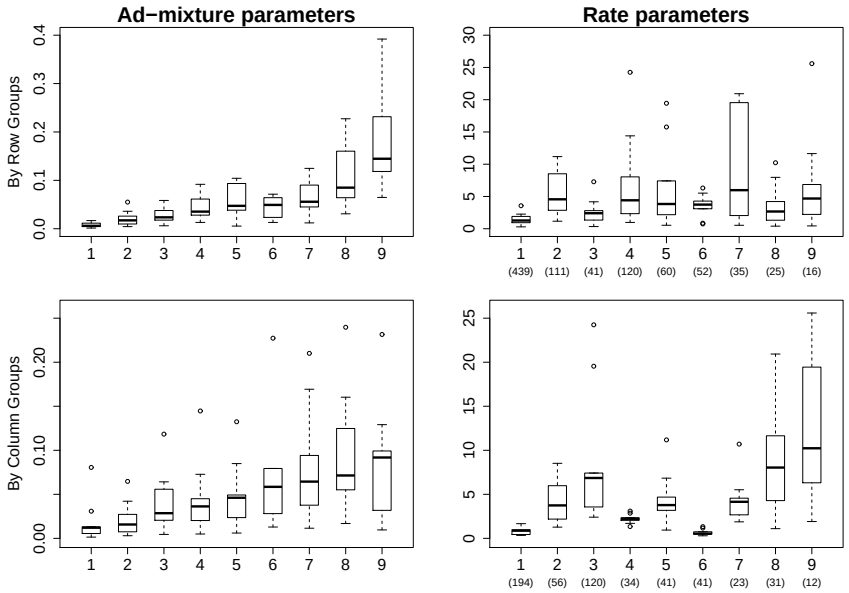


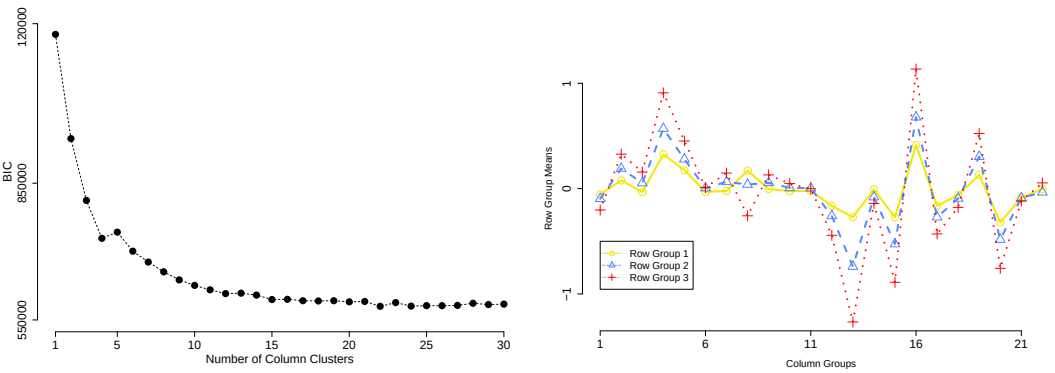
FIG 11. *Ad-mixture probabilities (left boxplots) and rate parameters (right boxplots) of the selected model on the UC-Irvine Forum dataset.*

fects have also been discussed on many biclustering applications including model-based collaborative filtering, network modelling, and microarray analysis. One additional advantage of covariate data is their ability to reduce the latent block search space. As demonstrated in the *MovieLens 100K* analysis, the best model has only 2 row and 3 column groups when both user and movie covariates are taken into account.

Another significant contribution of our study is the application of the variational generalized EM framework to the estimation of the biclustering models for large datasets. We have first extended the variational E step in the previous study (Vu, Hunter and Schweinberger, 2013) to handle both discrete and continuous responses. The comparison experiment between this new MM update and the popular fixed-point update clearly demonstrates its superior performance. New MM updates for the variational M step are also derived so that biclustering models with a large number of coefficients can be estimated. As illustrated in the *MovieLens 100K* analysis, these MM updates should be able to handle biclustering models with thousands of covariate coefficients. Our main explanation for the success of these variational generalized EM algorithms is their ability of separating parameters and weakening the dependence of corresponding updates.

We have also explored the idea of using the bootstrapping procedure in network analysis for biclustering model validation. Besides model selection criteria, these goodness of fit plots can provide us with a very powerful alternative tool for understanding and refining models. In the *UC-Irvine Forum* analysis, for example, the goodness of fit plot has helped to validate the zero-inflated model assumption. It is also important to note that this bootstrapping procedure can also be used to obtain standard errors of variational estimates as discussed in (Vu, Hunter and Schweinberger, 2013, Sections 5 and 7).

In future work, we plan to extend the proposed biclustering framework to temporal datasets where matrices are observed over time (Mankad and Michailidis, 2013). For example, gene expression can be measured on samples over their life cyle; or spending amounts on products can be



(a) BICs (b) Group Means

FIG 12. *(a) BICs of AGEMAP biclustering models where the number of row clusters is fixed at 3 and the number of column clusters is varied from 1 to 30. (b) Three row curves of mean parameters μ_{kl} are plotted across 22 column clusters.*

recorded for users every year. This temporal biclustering method can detect groups of genes more effectively since the timing resolution is expected to provide more fine-grained group structures. Regularization methods (Tibshirani, 1996) can also be applied to group-specific parameters to minimize changes within clusters while maximizing variations between them. Moreover, to scale the biclustering method to larger datasets, parallelization can be used to accelerate the computation. The separation of MM parameter updates lends themselves naturally to GPU programming (Zhou, Lange and Suchard, 2010).

The source code, written in `Java`, and data files used in Sections 5 and 6 will be made publicly available at the first author's website at `http://www.ms.unimelb.edu.au/~duyv`.

8. Acknowledgments. This research is supported by ARC Discovery Project DP120102902. We would also like to thank Professors Garry Robins and Pip Pattison for motivating us to work on this biclustering problem, and Professor Michael Schweinberger for his helpful comments on the paper. Experiments in this paper were run on the Edward computer cluster maintained by Unimelb Research Services HPC staff.

References.

AIROLDI, E., BLEI, D., FIENBERG, S. and XING, E. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research* **9** 1981–2014.

ARABIE, P., BOORMAN, S. A. and LEVITT, P. R. (1978). Constructing blockmodels: How and why. *Journal of Mathematical Psychology* **17** 21 - 63.

BARNDORFF-NIELSEN, O. E. (1978). *Information and Exponential Families in Statistical Theory*. Wiley, New York.

BEER, M. A. and TAVAZOIE, S. (2004). Predicting gene expression from sequence. *Cell* **117** 185–198.

BORGATTI, S. P. and EVERETT, M. G. (1997). Network analysis of 2-mode data. *Social Networks* **19** 243 - 269.

CHENG, Y. and CHURCH, G. M. (2000). Biclustering of expression data. *Proceedings of International Conference on Intelligent Systems for Molecular Biology* **8** 93–103.

CHO, E., MYERS, S. A. and LESKOVEC, J. (2011). Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. KDD '11* 1082–1090. ACM, New York, NY, USA.

DAUDIN, J. J., PICARD, F. and ROBIN, S. (2008). A mixture model for random graphs. *Statistics and Computing* **18** 173–183.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society Ser. B* **39**.

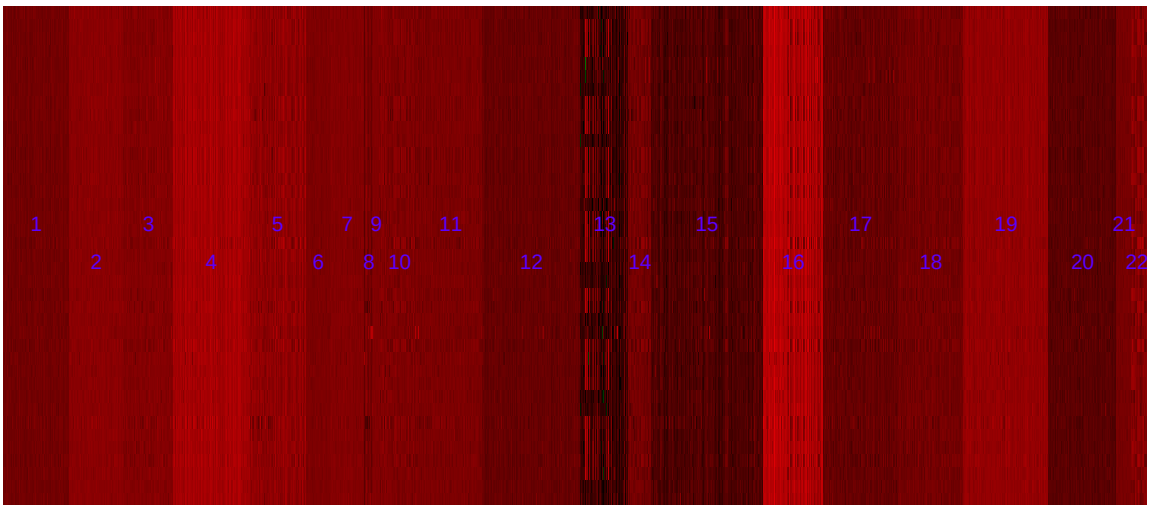
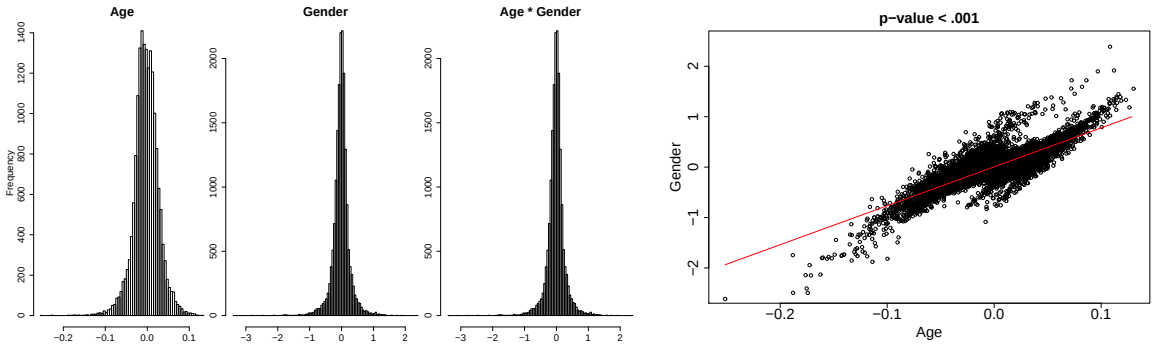


FIG 13. *The gene expression matrix AGEMAP is re-arranged using group assigments from the best biclustering model. Column group numbers are plotted at the central positions of their groups.*



(a) Column-varying coefficients (b) Coefficient correlation

FIG 14. *(a) Column-varying coefficients of row covariates of mouse age and gender, and their interaction from the best biclustering model in the AGEMAP analysis. (b) The correlation plot between age and gender coefficient estimates.*

DOREIAN, P., BATAGELJ, V. and FERLIGOJ, A. (2004). Generalized blockmodeling of two-mode network data. *Social Networks* **26** 29 - 53.

FLYNN, C. J. and PERRY, P. O. (2012). Consistent Biclustering. *ArXiv e-prints*.

FREEMAN, L. (2003). Finding Social Groups: A Meta-Analysis of the Southern Women Data. In *Dynamic Social Network Modeling and Analysis* (R. BREIGER, C. CARLEY and P. PATTISON, eds.) 39–97. National Research Council. The National Academies Press, Washington, DC.

GOODMAN, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **61** 215-231.

GOVAERT, G. and NADIF, M. (2008). Block clustering with Bernoulli mixture models: Comparison of different approaches. *Computational Statistics and Data Analysis* **52** 3233–3245.

GREENE, W. H. (1994). Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models Working Papers No. 94-10, New York University, Leonard N. Stern School of Business, Department of Economics.

HANISCH, D., ZIEN, A., ZIMMER, R. and LENGAUER, T. (2002). Co-clustering of biological networks and gene expression data. *Bioinformatics* **18** S145-S154.

HERLOCKER, J. L., KONSTAN, J. A., BORCHERS, A. and RIEDL, J. (1999). An algorithmic framework for performing

collaborative filtering. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. SIGIR '99* 230–237. ACM, New York, NY, USA.

Hunter, D. R., Goodreau, S. M. and Handcock, M. S. (2008). Goodness of Fit of Social Network Models. *Journal of the American Statistical Association* **103** 248-258.

Hunter, D. R. and Lange, K. (2004). A tutorial on MM algorithms. *The American Statistician* **58** 30–38.

Lambert, D. (1992). Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics* **34** 1–14.

Lange, K. (2010). *Numerical Analysis for Statisticians*. Springer Verlag Gmbh.

Lazarsfeld, P. F. and Neil, H. W. (1968). *Latent Structure Analysis*. Houghton Mifflin, Boston.

Lazzeroni, L. and Owen, A. (2002). Plaid models for gene expression data. *Statistica Sinica* **12** 61–86.

Madeira, S. C. and Oliveira, A. L. (2004). Biclustering Algorithms for Biological Data Analysis: A Survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **1** 24–45.

Mankad, S. and Michailidis, G. (2013). Biclustering Three-Dimensional Data Arrays with Plaid Models. *Technical report, Department of Statistics, University of Michigan*.

Mariadassou, M., Robin, S. and Vacher, C. (2010). Uncovering latent structure in valued graphs: A variational approach. *Annals of Applied Statistics* **4** 715–742.

Mirkin, B. (1996). *Mathematical classification and clustering*. Kluwer Academic Press.

Neal, R. M. and Hinton, G. E. (1993). A new view of the EM algorithm that justifies incremental and other variants. In *Learning in Graphical Models* 355–368. Kluwer Academic Publishers.

Opsahl, T. (2013). Triadic closure in two-mode networks: Redefining the global and local clustering coefficients. *Social Networks* **35** 159 - 167. Special Issue on Advances in Two-mode Social Networks.

Perry, P. O. and Owen, A. B. (2010). A Rotation Test to Verify Latent Structure. *J. Mach. Learn. Res.* **11** 603–624.

Salter-Townshend, M. and Murphy, T. B. (2013). Variational Bayesian inference for the latent position cluster model for network data. *Computational Statistics & Data Analysis* **57** 661–671.

Shan, H. and Banerjee, A. (2008). Bayesian Co-clustering. In *Data Mining, 2008. ICDM '08. Eighth IEEE International Conference on* 530-539.

Stefanov, S. M. (2004). Convex quadratic minimization subject to a linear constraint and box constraints. *Appl Math Res Express* **2004** 17-42.

Su, X. and Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques. *Advance in Artificial Intelligence* **2009** 4:2–4:2.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society (Series B)* **58** 267-288.

Vu, D. Q., Hunter, D. R. and Schweinberger, M. (2013). Model-based clustering of large networks. *Annals of Applied Statistics* **7** 1010–1039.

Wainwright, M. J. and Jordan, M. I. (2008). *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers Inc., Hanover, MA, USA.

Wang, P., Pattison, P. and Robins, G. (2013). Exponential random graph model specifications for bipartite networks - A dependence hierarchy. *Social Networks* **35** 211 - 222. Special Issue on Advances in Two-mode Social Networks.

Wang, Y., Print, C. and Crampin, E. (2013). Biclustering reveals breast cancer tumour subgroups with common clinical features and improves prediction of disease recurrence. *BMC Genomics* **14** 102.

Wang, P., Sharpe, K., Robins, G. L. and Pattison, P. E. (2009). Exponential random graph models for affiliation networks. *Social Networks* **31** 12 - 25.

Zahn, J. M., Poosala, S., Owen, A. B., Ingram, D. K., Lustig, A., Carter, A., Weeraratna, A. T., Taub, D. D., Gorospe, M., Mazan-Mamczarz, K., Lakatta, E. G., Boheler, K. R., Xu, X., Mattson, M. P., Falco, G., Ko, M. S. H., Schlessinger, D., Firman, J., Kummerfeld, S. K., Wood, W. H., Zonderman, A. B., Kim, S. K. and Becker, K. G. (2007). AGEMAP: A Gene Expression Database for Aging in Mice. *PLOS Genetics*.

Zhou, H. and Lange, K. (2009). Rating Movies and Rating the Raters Who Rate Them. *The American Statistician* **63** 297-307.

Zhou, H., Lange, K. and Suchard, M. A. (2010). Graphics Processing Units and High-Dimensional Optimization. *Statistical Science* **25** 311-324.

Department of Mathematics and Statistics
The University of Melbourne
Parkville, VIC, 3010
Australia
E-mail: duy.vu@unimelb.edu.au

Department of Mathematics and Statistics
The University of Melbourne
Parkville, VIC, 3010
Australia
E-mail: murray.aitkin@unimelb.edu.au