The Annals of Applied Statistics
2022, Vol. 16, No. 4, 2183–2200
https://doi.org/10.1214/21-AOAS1582

ESTIMATING HETEROGENEOUS GENE REGULATORY NETWORKS FROM ZERO-INFLATED SINGLE-CELL EXPRESSION DATA

BY QIUYU WU[a] AND XIANGYU LUO[b]

Institute of Statistics and Big Data, Renmin University of China, [a]*w.qy@ruc.edu.cn,* [b]*xiangyuluo@ruc.edu.cn*

Inferring gene regulatory networks can elucidate how genes work cooperatively. The gene-gene collaboration information is often learned by Gaussian graphical models (GGM) that aim to identify whether the expression levels of any pair of genes are dependent, given other genes' expression values. One basic assumption that guarantees the validity of GGM is data normality, and this often holds for *bulk-level* expression data which aggregate biological signals from a collection of cells. However, fine-grained *cell-level* expression profiles collected in single-cell RNA-sequencing (scRNA-seq) reveal non-normality features—cellular heterogeneity and zero inflation. We propose a Bayesian latent mixture GGM to jointly estimate multiple gene regulatory networks accounting for the zero inflation and unknown heterogeneity of single-cell expression data. The proposed approach outperforms competing methods on synthetic data in terms of network structure and precision matrix estimation accuracy and provides biological insights when applied to two real-world scRNA-seq datasets. An R package implementing the proposed model is available on GitHub https://github.com/WgitU/BLGGM.

1. Introduction. Genes are not independent workers but collaborate with each other to regulate associated biological processes. Elucidating gene regulatory networks allows us to get insights into underlying molecular mechanisms related to disease development, aging, and health (Chatterjee et al. (2016), Yang et al. (2015)). In Gaussian graphical models (GGM), the gene regulatory networks are encoded in a Gaussian graph, where nodes represent genes and edges capture the conditional dependence of expression levels of corresponding genes. Mathematically, let $(\theta_1, \theta_2, \ldots, \theta_p)$ be a random vector following a p-dimensional Gaussian distribution with mean vector $\boldsymbol{\mu}$ and precision matrix $\boldsymbol{\Omega}$, where θ_j represents the expression level of gene j for $1 \le j \le p$. Accordingly, the Gaussian graph is totally delineated by elements in $\boldsymbol{\Omega}$: there is no edge between nodes j_1 and j_2 ($j_1 \ne j_2$) in the Gaussian graph (i.e., θ_{j_1} and θ_{j_2} are conditionally independent) if and only if $\boldsymbol{\Omega}_{j_1 j_2} = 0$ (Dempster (1972), Yuan and Lin (2007), Friedman, Hastie and Tibshirani (2008)).

GGM requires normality of observed data which is a basic assumption to correctly recover gene regulatory networks. Fortunately, the normality often holds for *bulk-level* gene expression data, which are aggregate signals over all cells in a sample (Pratapa et al. (2020)), based on central limit theorem. With the fast development and increasing popularity of single-cell sequencing technology nowadays, such as single-cell RNA-sequencing (scRNA-seq), *cell-level* expression data become more common to researchers. However, cell-level expressions are different from bulk-level expressions in two ways. First, single-cell data are zero inflated, owing to the fact that there is a significantly smaller amount of mRNA molecules in one single cell than those in a bulk-level sample so that low cell-level expressions of some genes tend to be missed, resulting in zero values. Zero inflation is also called dropout, so we use the two exchangeable terms throughout the paper. Second, single-cell data capture cellular heterogeneity, and the distribution of heterogeneous expression values often exhibits

Received January 2021; revised August 2021.
Key words and phrases. Bayesian analysis, Gaussian graphical models, heterogeneity, nonignorable dropout, spike-slab prior.

multi-modality. Therefore, considering the increasing deposition of single-cell data in public databases (Edgar, Domrachev and Lash (2002), Rozenblatt-Rosen et al. (2017)), how to generalize GGM to account for zero inflation and cellular heterogeneity is a crucial problem in correctly estimating gene regulatory networks from single-cell expression data.

The application of GGM to recover gene regulatory networks from bulk-level expression has received attentions from both perspectives of frequentist (Yuan and Lin (2007), Friedman, Hastie and Tibshirani (2008)) and Bayesian (Dobra, Lenkoski and Rodriguez (2011), Wang (2012), Wang and Li (2012), Mohammadi and Wit (2015)). For example, glasso (Friedman, Hastie and Tibshirani (2008)) appended an L_1-norm penalty term of nondiagonals in the precision matrix $\mathbf{\Omega}$ to the likelihood of $\mathbf{\Omega}$ and then maximized penalized likelihood. In this way, some nondiagonals can be exactly estimated as zeros. Bayesian approaches aim to assign reasonable priors to $\mathbf{\Omega}$, which can induce sparsity of the precision matrices, for example, G-Wishart prior (Wang and Li (2012)), Bayesian glasso (Wang (2012)) and continuous spike-slab prior (Wang (2015)). Nevertheless, all of the methods are only applicable to homogeneous bulk-level expression data and hence might lead to problematic results when there exist sample heterogeneity.

When sample heterogeneity is known, in other words, we know the information about which class each sample comes from, several statistical methods extend GGM to jointly estimate multiple Gaussian graphs by borrowing strengths across classes. Frequentist approaches (Guo et al. (2011), Danaher, Wang and Witten (2014), Saegusa and Shojaie (2016), Ma and Michailidis (2016)) employed additional penalties that link elements of multiple precision matrices to induce similar sparsity structures across conditions. In the Bayesian paradigm, Peterson, Stingo and Vannucci (2015) used G-Wishart distributions and a hypergraph prior to connect multiple graphs. Lin et al. (2017) extended Bayesian GGM to analyze brain microarray data with spatial and temporal structures. Li, McCormick and Clark (2019) took advantage of a continuous spike-slab framework to realize Bayesian treatments of group and fused graphical lasso. Gan et al. (2019) discussed the theoretical underpinning of joint Bayesian estimation of multiple graphs using spike-slab lasso priors. When sample heterogeneity is not available, graphical models built upon mixture distributions (Rodríguez, Lenkoski and Dobra (2011), Gao et al. (2016), Luo and Wei (2018), Hao et al. (2018), Ren et al. (2021a)) were proposed to achieve simultaneous clustering and multiple graph estimations.

Despite the successful application of aforementioned graphical models to bulk-level expression, there are few statistical models that can deal with zero inflation in single-cell expression data. McDavid et al. (2019) proposed a Hurdle graphical model to account for zero-inflation of single-cell data. The Hurdle model turns out to be a mixture of degenerated Gaussian distributions and encodes conditional independences through three interaction matrices. Subsequently, they utilized the neighborhood selection technique to select edges for each node via penalized regression. Unfortunately, the Hurdle model does not consider the cellular heterogeneity among single-cell expressions. In addition, some computational biology methods (Aibar et al. (2017), Qiu et al. (2018)) also aim to reconstruct gene regulatory networks from homogeneous cells, but they rely on time-course expression data or expression data with estimated pseudo-time, while our work focuses on cross-sectional expression data.

To the best of our knowledge, there is a lack of statistical methods to estimate cell-type-specific gene regulatory networks from single-cell expression data that simultaneously consider zero inflation and cellular heterogeneity. Therefore, we developed a Bayesian latent Gaussian graph mixture model (BLGGM) to address the problem. The contributions of this paper are as follows: (1) The proposed model teases apart the cellular heterogeneity, using a model-based clustering strategy, and accounts for zero inflation through a nonignorable dropout mechanism. (2) We proved the model identifiability (up to clustering label permutation). (3) We inferred the model in the Bayesian paradigm thus enabling the quantification

of uncertainty of graph structures and gene-gene collaboration intensities. (4) The model has better performances in recovering gene regulatory networks than competing methods and provides valid biological results in the application to two real scRNA-seq datasets.

2. Method.

2.1. *Modeling cellular heterogeneity.* Suppose that there are n sequenced cells and each cell has G genes. We denote the true expression value of cell i on gene g by θ_{gi}. Considering the cellular heterogeneity, we assume that the n cells belong to K cell types (K is a positive integer) and utilize the model-based clustering (Fraley and Raftery (2002)) to model $\boldsymbol{\theta}_i = (\theta_{1i}, \theta_{2i}, \ldots, \theta_{Gi})$. Specifically, $\boldsymbol{\theta}_i$ is assumed to follow a mixture of K Gaussian distributions, $\boldsymbol{\theta}_i \sim \sum_{k=1}^{K} \pi_k \mathrm{N}(\boldsymbol{\mu}_k, \boldsymbol{\Omega}_k^{-1})$. Here, π_k represents the cell-type k proportion satisfying $0 < \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$, $\boldsymbol{\mu}_k$ is the mean expression profile of cell type k, and $\boldsymbol{\Omega}_k$ is cell type k's precision matrix. If we associate cell i with a cell-type indicator C_i that describes the cell type to which cell i belongs, then the mixture distribution for $\boldsymbol{\theta}_i$ is equivalent to

$$
\begin{aligned}
&\mathbb{P}(C_i = k) = \pi_k, \\
&\boldsymbol{\theta}_i | C_i = k \sim \mathrm{N}\big(\boldsymbol{\mu}_k, \boldsymbol{\Omega}_k^{-1}\big).
\end{aligned} \tag{1}
$$

These indicators $\{C_i : 1 \leq i \leq n\}$ are unknown and reflect the heterogeneity among cells.

2.2. *Modeling zero inflation.* The true expression level matrix $\{\theta_{gi} : 1 \leq g \leq G, 1 \leq i \leq n\}$ is not directly observed due to zero inflation (Risso et al. (2018)). In practice, we assume that the scRNA-seq raw count data are first normalized to account for library sizes (e.g., counts per the median library size of cells), resulting in the data matrix whose elements are continuous and nonnegative. Assuming X_{gi} is the actually observed expression of gene g in cell i after normalization, the relationship between X_{gi} and θ_{gi} is modeled as follows. Conditional on θ_{gi},

$$
X_{gi} = \begin{cases} 0 & \text{with probability } p(\theta_{gi}), \\ e^{\theta_{gi}} & \text{with probability } 1 - p(\theta_{gi}). \end{cases} \tag{2}
$$

$p(\theta_{gi})$ describes the probability that a dropout occurs on gene g in cell i and is defined as $\Phi(\lambda_{g0} + \lambda_{g1}\theta_{gi})$, $\lambda_{g1} < 0$. Φ is the cumulative distribution function of the standard normal distribution, and λ_{g0} and λ_{g1} depict the influence of θ_{gi} on the dropout event. The negativeness of λ_{g1} ensures that the stronger the signal of θ_{gi}, the less likely we observe a zero on this gene. Similar *nonignorable* dropout mechanism has been used in scRNA-seq analysis to model zero inflation (Song, Chan and Wei (2020)) (here, we borrow the term "nonignorable" from the missing data analysis as the zero inflation probability relies on the underlying value θ_{gi}).

Moreover, if the dropout event does not happen, the observed X_{gi} is assumed to be an exponential transformation of θ_{gi}. Given $C_i = k$, as θ_{gi} follows a Gaussian/normal distribution, $X_{gi} = e^{\theta_{gi}}$ comes from an asymmetric log-normal distribution by definition. The asymmetry feature has been observed in scRNA-seq data (Vieth et al. (2019)). In addition, log-normal-based distributions have been widely proposed to fit sequencing data in biological studies (Gallopin, Rau and Jaffrézic (2013), Zhang et al. (2015), Ntranos et al. (2019)). Hence, the usage of the exponential transformation ensures a well-grounded distribution for observed expression value X_{gi}.

2.3. *The unified model.* We subsequently combine equations (1) and (2) and obtain the following Bayesian latent Gaussian graph mixture (BLGGM) model:

$$
\begin{aligned}
&\mathbb{P}(C_i = k) = \pi_k,\\
&\boldsymbol{\theta}_i | C_i = k \sim \mathrm{N}\big(\boldsymbol{\mu}_k, \boldsymbol{\Omega}_k^{-1}\big),\\
&X_{gi} | \theta_{gi} = \begin{cases} 0 & \text{with probability } \Phi(\lambda_{g0} + \lambda_{g1}\theta_{gi}),\\ e^{\theta_{gi}} & \text{with probability } 1 - \Phi(\lambda_{g0} + \lambda_{g1}\theta_{gi}). \end{cases}
\end{aligned} \tag{3}
$$

In this model the unknown parameters are cell-type-k proportion π_k, mean expression profile $\boldsymbol{\mu}_k$, precision matrix $\boldsymbol{\Omega}_k$ which encode the gene regulatory networks for $1 \le k \le K$ and dropout-related coefficients $\boldsymbol{\lambda}_0 = (\lambda_{10}, \lambda_{20}, \ldots, \lambda_{G0})$ as well as $\boldsymbol{\lambda}_1 = (\lambda_{11}, \lambda_{21}, \ldots, \lambda_{G1})$. All elements of $\boldsymbol{\lambda}_1$ are negative. Given observed data $\mathbf{X} = \{X_{gi} : 1 \le g \le G, 1 \le i \le n\}$, the likelihood function of the parameters is

$$
\begin{aligned}
&L\big(\boldsymbol{\lambda}_0, \boldsymbol{\lambda}_1, (\pi_1, \boldsymbol{\mu}_1, \boldsymbol{\Omega}_1), \ldots, (\pi_K, \boldsymbol{\mu}_K, \boldsymbol{\Omega}_K) | \mathbf{X}\big)\\
&\quad = \prod_{i=1}^{n} \Bigg[\sum_{k=1}^{K} \pi_k \int \prod_{g=1}^{G} \big[\delta_0(X_{gi})\Phi(\lambda_{g0} + \lambda_{g1}\theta_{gi}) + \delta_{e^{\theta_{gi}}}(X_{gi})\big(1 - \Phi(\lambda_{g0} + \lambda_{g1}\theta_{gi})\big)\big]\\
&\qquad \cdot \mathrm{N}\big(\boldsymbol{\theta}_i; \boldsymbol{\mu}_k, \boldsymbol{\Omega}_k^{-1}\big)\, d\boldsymbol{\theta}_i \Bigg],
\end{aligned} \tag{4}
$$

where $\delta_a(\cdot)$ is the Dirac probability measure with point mass on a and $\mathrm{N}(\boldsymbol{\theta}_i; \boldsymbol{\mu}_k, \boldsymbol{\Omega}_k^{-1})$ is the multivariate normal density evaluated at $\boldsymbol{\theta}_i$ with mean $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k := \boldsymbol{\Omega}_k^{-1}$.

We proved the identifiability of model (3) up to label switching. The proof is based on results from Miao, Ding and Geng (2016) and can be found in Section S1 of the Supplementary Material (Wu and Luo (2022)).

THEOREM 2.1 (Identifiability of BLGGM). *If* $(\boldsymbol{\mu}_{k_1}, \boldsymbol{\Omega}_{k_1}) \ne (\boldsymbol{\mu}_{k_2}, \boldsymbol{\Omega}_{k_2})$ *for any* $k_1 \ne k_2$ *and* $L(\boldsymbol{\lambda}_0, \boldsymbol{\lambda}_1, (\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Omega}_k), k = 1, \ldots, K | \mathbf{X}) = L(\boldsymbol{\lambda}_0^*, \boldsymbol{\lambda}_1^*, (\pi_k^*, \boldsymbol{\mu}_k^*, \boldsymbol{\Omega}_k^*), k = 1, \ldots, K^* | \mathbf{X})$ *for any* $\mathbf{X}$, *then we have* $K = K^*$, $\boldsymbol{\lambda}_0 = \boldsymbol{\lambda}_0^*$, $\boldsymbol{\lambda}_1 = \boldsymbol{\lambda}_1^*$ *and* $(\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Omega}_k) = (\pi_{\rho(k)}^*, \boldsymbol{\mu}_{\rho(k)}^*, \boldsymbol{\Omega}_{\rho(k)}^*)$ *for some permutation* ρ *of* $\{1, 2, \ldots, K\}$.

2.4. *Local and global conditional independence.* Practically, the gene regulatory network is difficult to be constructed in a transcriptome-wide manner because this is a huge computational cost and, more importantly, we often need to filter out genes that do not satisfy some quality requirements during data preprocessing. Thus, we emphasize that, within one cell type, the interpretation for the precision matrix $\boldsymbol{\Omega}_{p\times p}$ on the selected p genes is usually different from that for the precision matrix $\boldsymbol{\Omega}_{p^*\times p^*}^*$ on the transcriptome-wide whole p^* genes ($p < p^*$). If we partition $\boldsymbol{\Omega}^*$ into submatrices $\begin{pmatrix} \boldsymbol{\Omega}_1^* & \boldsymbol{\Omega}_{12}^* \\ \boldsymbol{\Omega}_{21}^* & \boldsymbol{\Omega}_2^* \end{pmatrix}$, where the first diagonal block $\boldsymbol{\Omega}_1^*$ corresponds to the selected p genes, we then have $\boldsymbol{\Omega}_{p\times p} = \boldsymbol{\Omega}_1^* - \boldsymbol{\Omega}_{12}^* \boldsymbol{\Omega}_2^{*-1} \boldsymbol{\Omega}_{21}^*$.

We say that the matrix $\boldsymbol{\Omega}_{p\times p}$ encodes *local conditional independence*, while $\boldsymbol{\Omega}_1^*$ encodes *global conditional independence*. Specifically, a zero value of the (j_1, j_2) entry in $\boldsymbol{\Omega}_{p\times p}$ ($j_1 \ne j_2$) implies that the expressions of genes j_1 and j_2 are independent, given *other* $p - 2$ *selected genes* $\{1, 2, \ldots, p\} \setminus \{j_1, j_2\}$. In contrast, a zero value of the (j_1, j_2) entry in $\boldsymbol{\Omega}_{1,p\times p}^*$ implies that the expressions of genes j_1 and j_2 are independent, given *all other* $p^* - 2$ *genes* $\{1, 2, \ldots, p^*\} \setminus \{j_1, j_2\}$. Using the definitions, BLGGM aims to uncover the local conditional independence for the selected p genes, based on the estimates for $\boldsymbol{\Omega}_{p\times p}$, rather than the global conditional independence.

2.5. *Connection to Tobit models*. The Tobit model is a class of flexible statistical regression methods to mitigate the problem of zero inflation in the observations and has been widely used in statistics and econometrics (Amemiya (1984)). We justify that the proposed model BLGGM is actually a *smoothed* standard Tobit model in Section S2 of the Supplementary Material (Wu and Luo (2022)), so it can be placed in the Tobit context.

3. Bayesian inference.

3.1. *Prior specification*. We first focus on the prior assignment for precision matrices $\mathbf{\Omega}_k$'s. To decode gene network structures from $\mathbf{\Omega}_k$, we need a prior that can induce sparsity on the estimation of $\mathbf{\Omega}_k$. The sparsity-promotion property can be realized by three types of priors, G-Wishart prior (Wang and Li (2012)), Bayesian graphical lasso prior (Wang (2012)) and continuous spike-slab prior (Wang (2015)).

G-Wishart distribution uses a graph as a hyperparameter and constrains elements which correspond to empty edges to be zero in the precision matrix. However, the graph update strategy adds or deletes only one edge at a time and thus causes a slow exploration of the whole graph space (Wang and Li (2012), Wang (2015)). Under Bayesian graphical lasso prior, the posterior mode of $\mathbf{\Omega}_k$ is equivalent to the solution to the penalized likelihood maximization problem in glasso (Friedman, Hastie and Tibshirani (2008)). Since in the inference $\mathbf{\Omega}_k$ is often estimated by averaging continuous posterior samples rather than finding a posterior mode, Bayesian glasso cannot provide sparse structures in $\mathbf{\Omega}_k$. In contrast, the continuous spike-slab prior (Wang (2015)) enjoys computational convenience for its continuity feature and is able to induce sparse estimate by augmenting edge indicators. Hence, we adopted the continuous spike-slab prior for $\mathbf{\Omega}_k$'s.

Specifically, we introduce binary latent variables $\mathbf{Z}_k = (z_{k,jt} \in \{0, 1\} : 1 \le j \ne t \le G)$, and $z_{k,jt} = 1$ indicates there is an edge connecting nodes j and t in the cell-type-k gene regulatory network. When $z_{k,jt} = 1$, $\Omega_{k,jt}$ follows a normal distribution with a large variance $\mathrm{N}(0, v_1^2)$, corresponding to the dispersed slab component. When $z_{k,jt} = 0$, $\Omega_{k,jt}$ is from a normal distribution with a lower variance $\mathrm{N}(0, v_0^2)$, corresponding to the concentrated spike part. We assign exponential distributions with rate $\alpha/2$ to diagonals $\Omega_{k,jj}$ $(1 \le j \le G)$. The continuous spike-slab prior is then represented by

$$p(\mathbf{\Omega}_k|\mathbf{Z}_k, v_0, v_1, \alpha) = C(\mathbf{Z}_k, v_0, v_1, \alpha)^{-1} \prod_{j<t} \mathrm{N}\big(\Omega_{k,jt}; 0, v_{z_{k,jt}}^2\big) \cdot \prod_j \mathrm{Exp}(\Omega_{k,jj}; \alpha/2) \cdot \mathbb{I}\big(\mathbf{\Omega}_k \in M^+\big),$$

$$p(\mathbf{Z}_k|v_0, v_1, \xi, \alpha) = C(v_0, v_1, \xi, \alpha)^{-1} C(\mathbf{Z}_k, v_0, v_1, \alpha) \cdot \prod_{j<t} \big(\xi^{z_{k,jt}} (1-\xi)^{1-z_{k,jt}}\big),$$

where terms $C(\mathbf{Z}_k, v_0, v_1, \alpha)$ and $C(v_0, v_1, \xi, \alpha)$ are normalizing constants with tuning parameters v_0, v_1, ξ, α, and $\mathbb{I}(\mathbf{\Omega}_k \in M^+)$ means that $\mathbf{\Omega}_k$ must be in the cone of positive definite matrices.

Next, we specify the priors for other parameters in the proposed model. The prior of cell-type-k expression mean on gene g μ_{gk} is set as a normal distribution $\mathrm{N}(\eta_\mu, \tau_\mu^2)$. The prior of cell-type proportion $(\pi_1, \ldots, \pi_K)$ is a Dirichlet distribution $\mathrm{Dir}(\gamma_1, \ldots, \gamma_K)$. Zero-inflation-related parameters λ_{g0} and λ_{g1} are given weakly informative priors $\mathrm{N}(\eta_{\lambda_0}, \tau_{\lambda_0}^2)$ and $\mathrm{N}(\eta_{\lambda_1}, \tau_{\lambda_1}^2)\mathbb{I}(\lambda_{g1} < 0)$, respectively.

3.2. *Bayesian posterior inference*. The observed-data likelihood (4) is intractable, as the integration with respect to $\boldsymbol{\theta}_i$ has no explicit form. We thus take a data augmentation technique (Tanner and Wong (1987)) by involving the latent random variables $\mathbf{C} = (C_1, \ldots, C_n)$

and $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_i : 1 \le i \le n\}$ to form the posterior distribution of both unknown parameters and latent variables, which removes the integration and is more friendly to performing sampling,

$$
\begin{aligned}
&p(\boldsymbol{\lambda}_0, \boldsymbol{\lambda}_1, \mathbf{C}, \boldsymbol{\Theta}, (\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Omega}_k, \mathbf{Z}_k), k = 1, \ldots, K | \mathbf{X}) \\
&\quad \propto \prod_{i=1}^{n} \prod_{k=1}^{K} \Bigg[\pi_k \prod_{g=1}^{G} \big[\delta_0(X_{gi}) \Phi(\lambda_{g0} + \lambda_{g1}\theta_{gi}) + \delta_{e^{\theta_{gi}}}(X_{gi}) \big(1 - \Phi(\lambda_{g0} + \lambda_{g1}\theta_{gi})\big) \big] \\
&\qquad \cdot \mathrm{N}\big(\boldsymbol{\theta}_i; \boldsymbol{\mu}_k, \boldsymbol{\Omega}_k^{-1}\big) \Bigg]^{\mathbb{I}(C_i = k)} \\
&\qquad \cdot \prod_{k=1}^{K} p(\boldsymbol{\Omega}_k | \mathbf{Z}_k, v_0, v_1, \alpha) p(\mathbf{Z}_k | v_0, v_1, \xi, \alpha) \prod_{g=1}^{G} \mathrm{N}\big(\mu_{gk}; \eta_\mu, \tau_\mu^2\big) \\
&\qquad \cdot \mathrm{Dir}(\pi_1, \ldots, \pi_K | \gamma_1, \ldots, \gamma_K) \cdot \prod_{g=1}^{G} \mathrm{N}\big(\lambda_{g0}; \eta_{\lambda_0}, \tau_{\lambda_0}^2\big) \mathrm{N}\big(\lambda_{g1}; \eta_{\lambda_1}, \tau_{\lambda_1}^2\big) \mathbb{I}(\lambda_{g1} < 0).
\end{aligned}
$$

Subsequently, we derive full conditional distributions for each parameter and latent variable and perform Gibbs sampler (Geman and Geman (1984), Gelman et al. (2013)). However, updates for $\boldsymbol{\Theta}$, $\boldsymbol{\lambda}_0$ and $\boldsymbol{\lambda}_1$ in Gibbs sampler are not of standard form, and traditional solutions, such as random-walk Metropolis–Hastings step (Metropolis et al. (1953)), suffer from exploration inefficiency and high correlations between nearby posterior samples. Thus, we resort to the Hamiltonian dynamic to obtain proposals that can be far from current position using gradient information which is often more efficient and significantly reduces between-sample correlation (Neal (2011)). Therefore, the proposed hybrid sampling scheme alternates between Gibbs sampler and Hamiltonian Monte Carlo (HMC), and it proceeds as follows ("$-$" means "given all other variables"):

1. (HMC) Update missing variable θ_{gi} for which X_{gi} equals zero.

Let $\boldsymbol{\theta}_{i,\mathrm{mis}}$ be the vector of $\{\theta_{gi} : g \in \{g : X_{gi} = 0\}\}$ and $\boldsymbol{\theta}_{i,\mathrm{obs}}$ be the vector of $\{\theta_{gi} : g \in \{g : X_{gi} > 0\}\}$. We then partition $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k = \boldsymbol{\Omega}_k^{-1}$ by the "mis" and "obs" parts, giving $\binom{\boldsymbol{\mu}_{k,\mathrm{obs}}}{\boldsymbol{\mu}_{k,\mathrm{mis}}}$ and $\left(\begin{smallmatrix} \Sigma_{k,\mathrm{obs}} & \Sigma_{k,12} \\ \Sigma_{k,21} & \Sigma_{k,\mathrm{mis}} \end{smallmatrix}\right)$, respectively. Given $C_i = k$, the conditional distribution of $\boldsymbol{\theta}_{i,\mathrm{mis}}$ is

$$
p(\boldsymbol{\theta}_{i,\mathrm{mis}} | -) = \mathrm{N}\big(\boldsymbol{\theta}_{i,\mathrm{mis}}; \boldsymbol{\mu}_k^*, \Sigma_k^*\big) \prod_{g: X_{gi} = 0} \Phi(\lambda_{g0} + \lambda_{g1}\theta_{gi}),
$$

where $\boldsymbol{\mu}_k^* = \boldsymbol{\mu}_{k,\mathrm{mis}} + \Sigma_{k,21}\Sigma_{k,\mathrm{obs}}^{-1}(\boldsymbol{\theta}_{i,\mathrm{obs}} - \boldsymbol{\mu}_{k,\mathrm{obs}})$ and $\Sigma_k^* = \Sigma_{k,\mathrm{mis}} - \Sigma_{k,21}\Sigma_{k,\mathrm{obs}}^{-1}\Sigma_{k,12}$.

2. (HMC) Update zero-inflation intensity parameters λ_{g0} and λ_{g1} from

$$
\begin{aligned}
p(\lambda_{g0}, \lambda_{g1} | -) \propto &\prod_{i: X_{gi} > 0} \big(1 - \Phi(\lambda_{g0} + \lambda_{g1}\theta_{gi})\big) \prod_{i: X_{gi} = 0} \Phi(\lambda_{g0} + \lambda_{g1}\theta_{gi}) \\
&\cdot \mathrm{N}\big(\lambda_{g0}; \eta_{\lambda_0}, \tau_{\lambda_0}^2\big) \cdot \mathrm{N}\big(\lambda_{g1}; \eta_{\lambda_1}, \tau_{\lambda_1}^2\big) \mathbb{I}(\lambda_{g1} < 0).
\end{aligned}
$$

3. (Standard Gibbs sampling) Update cell-type k expression mean profile $\boldsymbol{\mu}_k$ from the multivariate normal distribution with mean vector $(n_k\Sigma_k^{-1} + I/\tau_\mu^2)^{-1}(\Sigma_k^{-1}\sum_{i: C_i = k}\boldsymbol{\theta}_i + \eta_\mu/\tau_\mu^2)$ and covariance matrix $(n_k\Sigma_k^{-1} + I/\tau_\mu^2)^{-1}$, where n_k is the current number of cells in cell type k and I is a $G \times G$ identity matrix.

4. (Standard Gibbs sampling) Update precision matrices $\boldsymbol{\Omega}_k$'s and edge indicators $\boldsymbol{Z}_k$'s.

Following Wang (2015), we update $\boldsymbol{\Omega}_k$ column by column. Without loss of generality, we focus on the last column. Let $\boldsymbol{V}_k = (v_{z_{k,jt}}^2)$ be a $G \times G$ symmetric matrix with diagonals

being zeros. Partition $\boldsymbol{\Omega}_k$, $\boldsymbol{S}_k = \sum_{i:C_i=k}(\boldsymbol{\theta}_i - \boldsymbol{\mu}_k)(\boldsymbol{\theta}_i - \boldsymbol{\mu}_k)^\top$ and $\boldsymbol{V}_k$ as follows:

$$\boldsymbol{\Omega}_k = \begin{pmatrix} \boldsymbol{\Omega}_{k,11}, \boldsymbol{\Omega}_{k,12} \\ \boldsymbol{\Omega}_{k,12}^\top, \Omega_{k,22} \end{pmatrix}, \qquad \boldsymbol{S}_k = \begin{pmatrix} \boldsymbol{S}_{k,11}, \boldsymbol{s}_{k,12} \\ \boldsymbol{s}_{k,12}^\top, s_{k,22} \end{pmatrix}, \qquad \boldsymbol{V}_k = \begin{pmatrix} \boldsymbol{V}_{k,11}, \boldsymbol{v}_{k,12} \\ \boldsymbol{v}_{k,12}^\top, 0 \end{pmatrix}.$$

Then, sample $(\boldsymbol{\Omega}_{k,12}|-) \sim \mathrm{N}(-\boldsymbol{C}\boldsymbol{s}_{k,12}, \boldsymbol{C})$ and $(\Omega_{k,22} - \boldsymbol{\Omega}_{k,12}^\top \boldsymbol{\Omega}_{k,11}^{-1}\boldsymbol{\Omega}_{k,12}|-) \sim \Gamma(\frac{n_k}{2} + 1, \frac{s_{22}+\alpha}{2})$, where $\boldsymbol{C} = \{(s_{k,22}+\alpha)\boldsymbol{\Omega}_{k,11}^{-1} + \mathrm{diag}(\mathbf{v}_{k,12})^{-1}\}^{-1}$.

Subsequently, update latent variables $\boldsymbol{Z}_k$ independently from Bernoulli distributions with success probability $\mathbb{P}(z_{k,jt} = 1|-) = \frac{\mathrm{N}(\Omega_{k,jt};0,v_1^2)\xi}{\mathrm{N}(\Omega_{k,jt};0,v_1^2)\xi + \mathrm{N}(\Omega_{k,jt};0,v_0^2)(1-\xi)}$.

5. (Standard Gibbs sampling) Update cell-type indicators C_i for cell $i = 1, \ldots, n$ from the distribution $\mathbb{P}(C_i = k|-) = \pi_k \mathrm{N}(\boldsymbol{\theta}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)/\sum_{j=1}^K \pi_j \mathrm{N}(\boldsymbol{\theta}_i; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, $k = 1, \ldots, K$.

6. (Standard Gibbs sampling) Update cell-type proportions $(\pi_1, \ldots, \pi_K)$ from the Dirichlet distribution $\mathrm{Dir}(n_1 + \gamma_1, \ldots, n_k + \gamma_K)$.

Details regarding the implementation of HMC using leapfrog steps are listed in Section S3 of the Supplementary Material (Wu and Luo (2022)).

3.3. *Graph structure inference.* We define the posterior probability of inclusion (PPI) for edge (j, t) in cell type k as $\mathrm{PPI}_{k,jt} = \mathbb{P}(z_{k,jt} = 1|\mathbf{X})$, and it is approximated based on posterior samples of $z_{k,jt}$ through $\sum_{\ell=1}^L \mathbb{I}(z_{k,jt}^{(\ell)} = 1)/L$ for $j \neq t$, where L is the number of posterior samples after the burn-in period. Subsequently, we infer the graph structures by controlling the expected Bayesian false discovery rate which is defined as follows (Newton et al. (2004), Peterson, Stingo and Vannucci (2015)):

$$\mathrm{FDR}(\kappa) = \frac{\sum_{k=1}^K \sum_{1 \le j < t \le G} \xi_{k,jt} \mathbb{I}(\xi_{k,jt} \le \kappa)}{\sum_{k=1}^K \sum_{1 \le j < t \le G} \mathbb{I}(\xi_{k,jt} \le \kappa)},$$

where $\xi_{k,jt} = 1 - \mathrm{PPI}_{k,jt}$. Generally, we choose an appropriate κ such that the Bayesian FDR is less than a threshold, such as 0.05. Peterson, Stingo and Vannucci (2015) claim that $\kappa = 0.5$ often results in a reasonable Bayesian FDR, so we follow their rule by cutting the PPI at $\kappa = 0.5$. Hence, $z_{k,jt}$ is estimated to be 1 if $\xi_{k,jt} \le \kappa$ and 0 otherwise. Actually, our simulation studies also justify that the FDR can be well controlled using this fixed $\kappa = 0.5$.

3.4. *The choice of the cell-type number.* We recommend the use of a modified Bayesian information criterion (penalized BIC) considering model sparsity (Pan and Shen (2007)) to find the optimal cell-type number K. The formula of pBIC in our case is

$$\mathrm{pBIC}(K) = -2\log\big(L\big(\hat{\boldsymbol{\lambda}}_0, \hat{\boldsymbol{\lambda}}_1, (\hat{\pi}_k, \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Omega}}_k), k = 1, \ldots, K|\mathbf{X}\big)\big) + \log(n)(d - d_0).$$

d is the number of parameters in the model and equals $K - 1 + G(2 + K + (G + 1)K/2)$. d_0 is the number of zero entries in the estimated precision matrices and equals $\sum_{k=1}^K \sum_{j=1}^{G-1} \sum_{t=j+1}^G \mathbb{I}(\hat{z}_{k,jt} = 0)$, and $\hat{z}_{k,jt}$ is the estimate of $z_{k,jt}$. $\hat{\boldsymbol{\lambda}}_0, \hat{\boldsymbol{\lambda}}_1, (\hat{\pi}_k, \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Omega}}_k), k = 1, \ldots, K$ are the posterior means of corresponding parameters. The details to calculate the pBIC value is given in Section S4 of the Supplementary Material (Wu and Luo (2022)).

3.5. *Detection of differential partial correlations.* An edge (j, t) is called *differential partial correlations* between cell types 1 and 2 if this edge is present in the two cell types but has partial correlations with opposite signs. Using posterior samples, we can easily estimate the probability of differential partial correlations, $\mathbb{P}(\Omega_{1,jt} > 0, \Omega_{2,jt} < 0|z_{1,jt} = z_{2,jt} = 1, \mathbf{X}) + \mathbb{P}(\Omega_{1,jt} < 0, \Omega_{2,jt} > 0|z_{1,jt} = z_{2,jt} = 1, \mathbf{X})$.

3.6. *Tied precision matrices*. Given the cell-type number K ($K \geq 2$), we may be curious about whether the precision matrix contributes to the cell heterogeneity. To that end, we calculate the pBIC value when $\mathbf{\Omega}_1 = \mathbf{\Omega}_2 = \cdots = \mathbf{\Omega}_K$, where d becomes $K - 1 + G(2 + K + (G + 1)/2)$, and compare it to original pBIC(K). If the latter is smaller, the gene regulation networks play a role in differentiating cells.

4. Simulation study. We generated data following model (3) with $K = 3$ cell types, $n = 3000$ cells and $G = 100$ genes. The first 30% genes were marker genes that exhibit differential expression levels in at least two cell types, and each of the rest genes has the same expression mean across cell types. Specific values of mean expression profiles were presented in Section S5 of the Supplementary Material (Wu and Luo (2022)). Cell-type proportion vector (π_1, π_2, π_3) was set to be $(0.4, 0.3, 0.3)$. Dropout-related coefficients λ_{g0} and λ_{g1} were sampled from $\mathrm{N}(1, 0.1^2)$ and $\mathrm{N}(-1, 0.1^2)$, respectively, for each gene g. The interquartile range (IQR) of cellwise zero proportions is [19.0%, 25.0%] with median 22.0% and maximum 38.0% in the simulated data.

Next, we specified the precision matrices $\mathbf{\Omega}_k$ for $1 \leq k \leq K$. Each $\mathbf{\Omega}_k$ was set as a block diagonal matrix, and every block is one of the following four modules:

1. Dense module $\mathbf{M}_d$: a 10 by 10 matrix with elements $M_{d,jj} = 2$; $M_{d,jt} = 0.7$ for $0 < |j - t| \leq 5$ and $M_{d,jt} = 0$, otherwise.
2. Circle module $\mathbf{M}_c$: a 10 by 10 matrix with elements $M_{c,jj} = 2$; $M_{c,jt} = 0.9$ for $|j - t| = 1$; $M_{c,1,10} = M_{c,10,1} = 0.9$ and $M_{d,jt} = 0$, otherwise.
3. Star module $\mathbf{M}_s$: a 10 by 10 matrix where node 1 is the central role that connects to all other nodes: $M_{s,jj} = 2$, $M_{s,1j} = M_{s,j1} = 0.6$ for $2 \leq j \leq 10$ and $M_{s,jt} = 0$, otherwise.
4. Partially negative dense module $\mathbf{M}_{nd}$: a 10 by 10 matrix with elements $M_{nd,jj} = 2$; $M_{nd,jt} = M_{nd,tj} = -0.6$ for $0 < j - t \leq 5$ and $t = 1, 2$; $M_{nd,jt} = M_{nd,tj} = 0.6$ for $0 < j - t \leq 5$ and $t = 3, 4, 5$ and $M_{nd,jt} = 0$, otherwise.

$\mathbf{\Omega}_1$ consists of 10 modules: the first three are circle modules, the next is a partially negative dense module and the last six blocks are dense modules. If we denote this precision matrix type by $[3c, 1nd, 6d]$, where "c" means circle, "nd" represents partially negative dense and "d" is dense, then the types of $\mathbf{\Omega}_2$ and $\mathbf{\Omega}_3$ are $[3d, 7d]$ and $[3s, 7d]$ ("s" means star), respectively. Heatmaps of the three precision matrices on the first 50 genes were displayed in Figure 1(a). We used 50 genes for a good visualization, and the figure on whole 100 genes is shown in Figure S1 of the Supplementary Materials (Wu and Luo (2022)).

We set hyperparameters $v_0 = 0.02$, $v_1 = 1$, $\xi = 2/(G-1)$ and $\alpha = 1$, as suggested in Wang (2015). In the Bayesian inference procedure we performed 10,000 iterations (time cost: 25.94 mins using 24 cores), and samples in the last 5000 iterations were kept. Markov chain has reached stationary (Figure S2 of the Supplementary Material (Wu and Luo (2022))). Continuous parameters $\boldsymbol{\lambda}_0$, $\boldsymbol{\lambda}_1$, $\boldsymbol{\pi}$, $\boldsymbol{\mu}_k$ and $\mathbf{\Omega}_k$ were estimated by posterior means. The estimates of $\mathbf{\Omega}_k$'s were shown in Figure 1(b). Posterior inclusive probabilities for binary indicators $z_{k,jt}$ quantify the certainty that there is a connection between genes i and j in cell type k (Figure 1(c)). In addition, we calculated pBIC values for K from two to six. According to Figure 2(a), pBIC attains minimum when $K = 3$ which is the truth. Given $K = 3$, pBIC for tied precision matrices is 814,830.6 while it is 800,817.2 for heterogeneous precision matrices which is also consistent with the truth.

Figure 3 shows the network structure estimates in cell type 1 from the proposed model and competing approaches including BDgraph (Mohammadi and Wit (2015), Mohammadi and Wit (2019)), GGMPF (Ren et al. (2021a, 2021b)), glasso (Friedman, Hastie and Tibshirani (2008)), HurdleNormal (McDavid et al. (2019)) and ppcor (Kim (2015)). For the network recovery performance in cell types 2 and 3, please refer to Figure S3 of the Supplementary

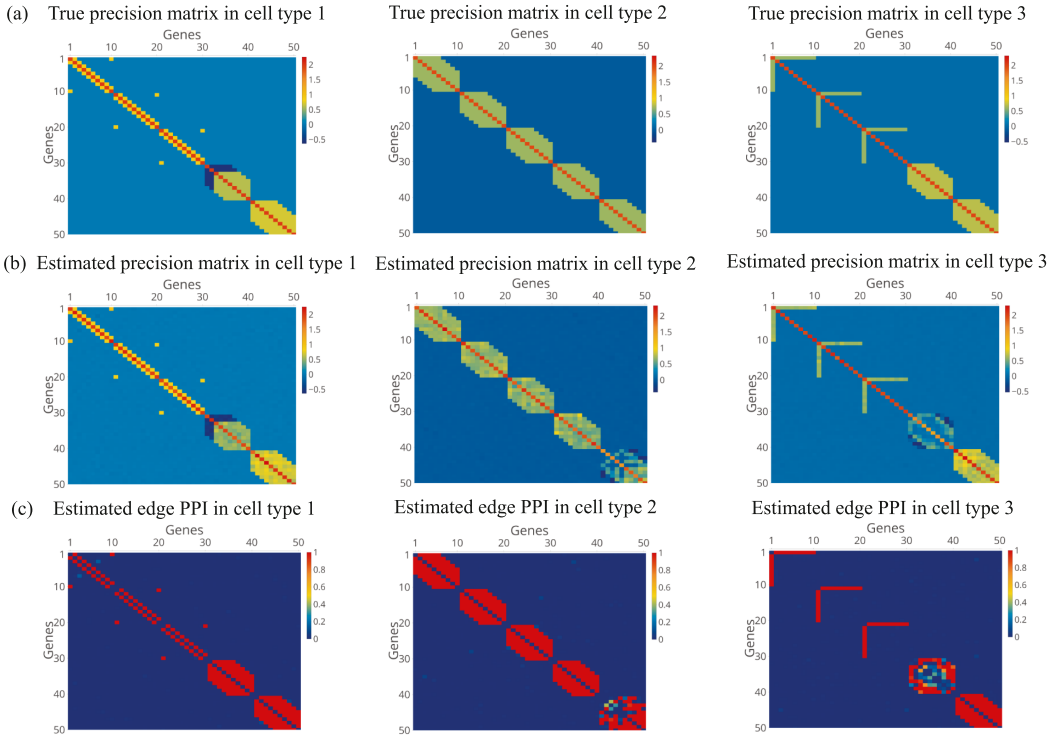

FIG. 1. *Comparisons between*: (*a*) *true precision matrices and* (*b*) *estimated precision matrices for each cell type on the first* 50 *genes*. (*c*) *Posterior probability of inclusion* (*PPI*) *for each edge*.

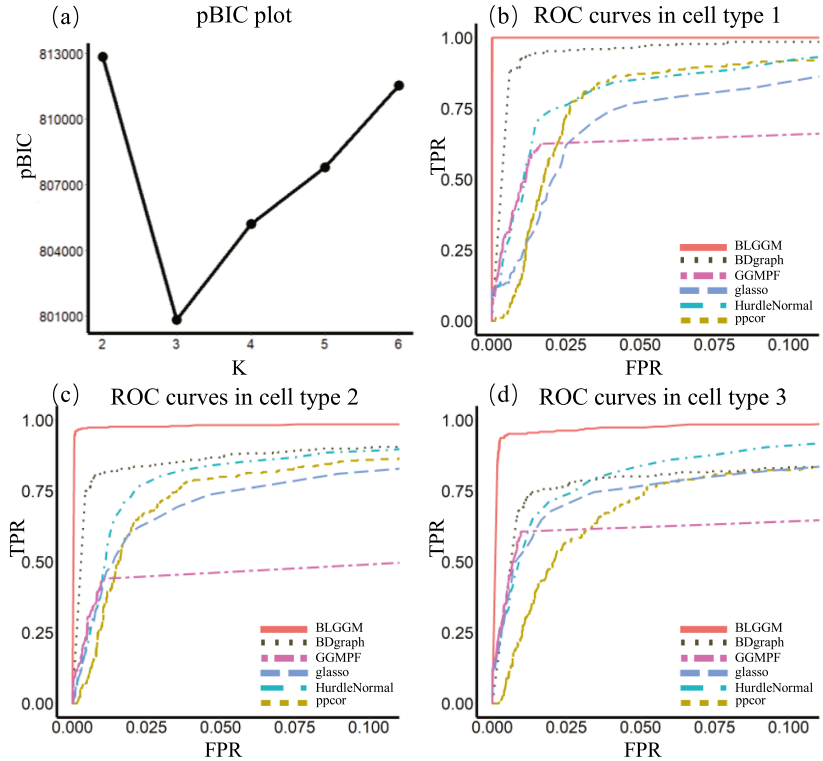

FIG. 2. *The results of the simulation*: (*a*) *The pBIC plot for K from* 2 *to* 6. (*b*) *ROC curves with FPR less than* 0.1 *in cell type* 1. (*c*) *ROC curves with FPR less than* 0.1 *in cell type* 2. (*d*) *ROC curves with FPR less than* 0.1 *in cell type* 3.

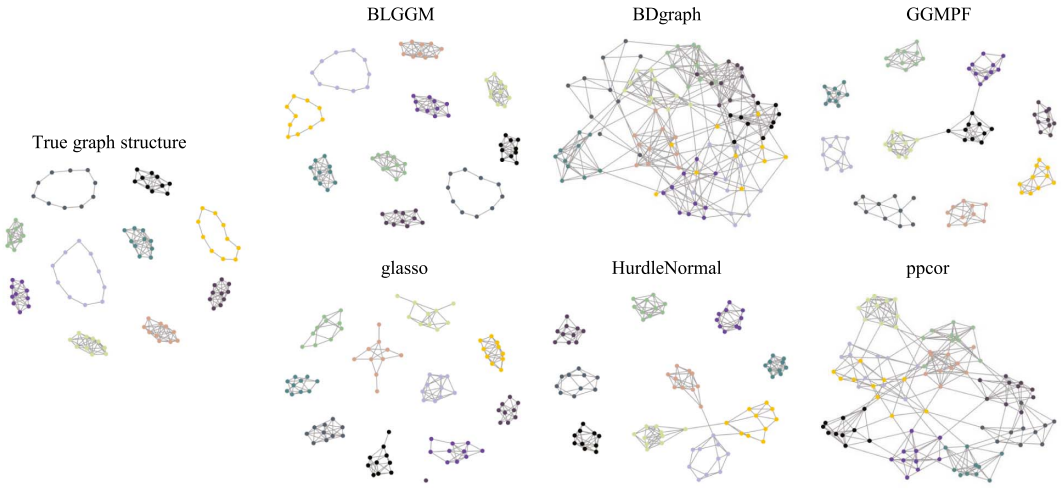

FIG. 3. *Performances of recovering gene regulatory network in cell type* 1.

Material (Wu and Luo (2022)). In terms of clustering accuracy, we computed the adjusted Rand index (ARI) (Hubert and Arabie (1985)) between BLGGM estimates and true cell-type labels, giving a perfect clustering (Table 1). Among competing approaches, only GGMPF is able to conduct clustering (mean ARI $= 0.94$), while others cannot automatically learn the cellular heterogeneity. Thus, we applied them in an oracle situation where the cell-type information is available. Table 1 provides comparisons in terms of edge-detection true positive rate (TPR), false positive rate (FPR), false discovery rate (FDR) and Frobenius norm (F-norm) between estimated precision matrices and the truth. Compared to other methods, the proposed model not only estimated network structures well (low FPR, FDR and high TPR) but also gave accurate estimates for elements in the precision matrices (low F-norm). Implementation details can be found in Section S6 of the Supplementary Material (Wu and Luo (2022)). Their ROC curves with FPR less than 0.1 are also provided in Figure 2(b)–(d).

Notice that, for the fourth module, some elements in the precision matrix of cell type 1 are negative, while they are positive in cell types 2 and 3 (Figure 1(a)). To detect edges

TABLE 1
Comparisons are based on ten replications in the simulation. Numbers in parentheses represent standard deviations

	Cell type	BLGGM	BDgraph	GGMPF	glasso	HurdleNormal	ppcor
TPR	1	0.94 (0.06)	0.87 (0.04)	0.57 (0.05)	0.50 (0.02)	0.70 (0.01)	0.74 (0.03)
	2	0.88 (0.05)	0.82 (0.03)	0.42 (0.03)	0.59 (0.02)	0.71 (0.02)	0.63 (0.03)
	3	0.85 (0.05)	0.78 (0.04)	0.58 (0.03)	0.56 (0.02)	0.77 (0.03)	0.54 (0.05)
FPR	1	0.00 (0.00)	0.02 (0.00)	0.02 (0.01)	0.02 (0.00)	0.02 (0.00)	0.03 (0.00)
	2	0.00 (0.00)	0.03 (0.00)	0.01 (0.00)	0.02 (0.00)	0.02 (0.00)	0.02 (0.00)
	3	0.00 (0.00)	0.03 (0.00)	0.01 (0.00)	0.02 (0.00)	0.04 (0.00)	0.03 (0.00)
FDR	1	0.02 (0.02)	0.33 (0.02)	0.31 (0.07)	0.39 (0.02)	0.31 (0.02)	0.38 (0.02)
	2	0.04 (0.02)	0.30 (0.02)	0.25 (0.01)	0.31 (0.01)	0.28 (0.02)	0.33 (0.02)
	3	0.06 (0.03)	0.40 (0.03)	0.24 (0.03)	0.32 (0.04)	0.48 (0.02)	0.46 (0.03)
F-norm	1	4.77 (2.02)	13.34 (0.53)	17.03 (0.06)	21.01 (0.13)	NA	NA
	2	7.84 (1.49)	14.60 (0.41)	19.58 (0.09)	22.09 (0.14)	NA	NA
	3	7.73 (1.23)	14.45 (0.43)	17.10 (0.15)	19.89 (0.13)	NA	NA
ARI		1.00 (0.00)	NA	0.94 (0.14)	NA	NA	NA

with differential partial correlations, using the cell type 1 as the reference, we calculated $P(\Omega_{1,ij} > 0, \Omega_{2,ij} < 0 | z_{1,ij} = z_{2,ij} = 1) + P(\Omega_{1,ij} < 0, \Omega_{2,ij} > 0 | z_{1,ij} = z_{2,ij} = 1)$ for any edge (i, j) that is present in both cell types 1 and 2. The same way was applied to identify differential partial correlations between cell types 1 and 3. Figure S4 of the Supplementary Material (Wu and Luo (2022)) reported the heatmap of the tail probabilities, and if we use 0.5 as a threshold, the underlying positions where partial correlations have different signs can be well recovered with TPR $= 1$, FPR $= 0$, FDR $= 0$ between cell types 1 and 2, and with TPR $= 0.6$, FPR $= 0$, FDR $= 0$ between cell types 1 and 3.

4.1. *Scale-free networks*. We used the function "sample_pa" in the Rpackage "igraph" (Csardi, Nepusz et al. (2006)) to generate three scale-free networks with 100 vertices and then simulated precision matrices with support on the network structures. Figure S5 of the Supplementary Material (Wu and Luo (2022)) displays the heatmaps of the three precision matrices. Table S1 and Figure S6 of the Supplementary Material (Wu and Luo (2022)) show that BLGGM still outperforms competing methods regarding the network structure recovery and precision matrix estimation.

4.2. *Sensitivity*: *Model misspecification*. For the nondropout part in equation (3), we let data X_{gi} be generated from a count-valued Poisson distribution with mean $e^{\theta_{gi}}$ rather than be equal to the continuous value $e^{\theta_{gi}}$. We then transformed X_{gi} into $\tilde{X}_{gi}$ by $\tilde{X}_{gi} = X_{gi}/\ell_i \cdot \text{median}_i \ell_i$, where ℓ_i is the library size of cell i. Subsequently, the transformed data matrix $\{\tilde{X}_{gi} : 1 \le g \le G, 1 \le i \le n\}$ was used as input of our method. The ROC curves for edge detection were drawn in Figure S7 of the Supplementary Material (Wu and Luo (2022)). Compared to the ideal case (the proposed model is accurate), our method did not lose much power while controlling false positive rate. Thus, our model is robust to the misspecified case.

4.3. *Sensitivity*: *Normalization strategies*. We acknowledge that different normalization approaches for high-throughput genomic data can lead to different performances in the downstream analysis. Hence, we investigated how they influence the results of edge detection. Two other commonly used approaches, count per million (CPM) and quantile normalization (QN) method, were chosen. Figure S8 of the Supplementary Material (Wu and Luo (2022)) shows the ROC curves of the three normalization strategies, respectively. It is observed that scaling using the median library size and CPM outperforms QN in our method, so we recommended that users had better choose scaling by median or CPM normalization when they apply the proposed method.

4.4. *Performance and computational cost with more genes*. We also showed the performances of BLGGM and competing methods on gene numbers $G = 200$, 300 and 400. We can observe in Figure S9 of the Supplementary Material (Wu and Luo (2022)) that, as the gene number grows, the computational cost of BLGGM increases quadratically. The computational speed is 0.15 seconds per iteration with 100 genes and attains 7.5 seconds per iteration with 400 genes. Thus, we suggest the users choose, at most, 400 genes when conducting posterior inference. Moreover, in terms of network estimation accuracy, Table 1 and Tables S2–S4 of the Supplementary Material (Wu and Luo (2022)) indicate that BLGGM uniformly outperforms competing methods when $G = 200$, 300 and 400.

4.5. *Performance with various zero levels*. We further adjusted the median zero proportions from 25% to 35% and 45% with 100 genes. We observe that, with the increasing zero proportions, the network structure estimation accuracy of all the approaches is decreasing

(Figure 2(b)–(d) and Figure S10 of the Supplementary Material (Wu and Luo (2022))). However, in most cases, BLGGM has better performances than competing approaches. For example, even with the median 45% zero proportion, our method can still achieve a relatively high power 0.56 with a controlled FDR 0.15 (Table S6 of the Supplementary Material (Wu and Luo (2022))). In addition, Table 1 displays the clustering results for BLGGM and GGMPF when the median zero proportion is 25%, both of the methods can cluster cells well. However, as the zero proportion increases to 35% and 45%, GGMPF's clustering is less accurate than BLGGM (Tables S5–S6 of the Supplementary Material (Wu and Luo (2022))). Therefore, overall, the proposed method outperforms GGMPF in terms of clustering accuracy, thanks to the ability of BLGGM to handle the zero inflation in the observed single-cell expression data.

5. Real application.

5.1. *Mouse hematopoietic stem and progenitor cell (HSPC) data.* HSPCs have the ability to produce mature blood cells and show heterogeneity of self-renewal potential (Morita, Ema and Nakauchi (2010)). Nestorowa et al. (2016) sequenced 1920 HSPCs from mice. In data preprocessing, we followed the quality control scheme in Nestorowa et al. (2016), resulting in 1656 cells, and then divided raw scRNA-seq counts for each cell by its size factor which is defined as the ratio of the cell's library size to the median of library sizes across all cells. The normalized expression values were the input data $\mathbf{X}$ of our proposed model. Our interest is to construct gene regulatory networks of 40 marker genes for $K = 4$ HSPC subtypes detected by Nestorowa et al. (2016). The IQR of cellwise zero proportions is [22.5%, 37.5%], and its maximum is 70.0%. In fact, we tried multiple choices of K, ranging from 2 to 6, the pBIC plot in Figure 4(a) justifies the usage of $K = 4$. Moreover, when the cell-type number is four, we obtained pBIC = 365,837.7 for equal precision matrices and pBIC = 348,657.3 for different precision matrices which indicates that the gene-gene relationships indeed play a role in cell heterogeneity.

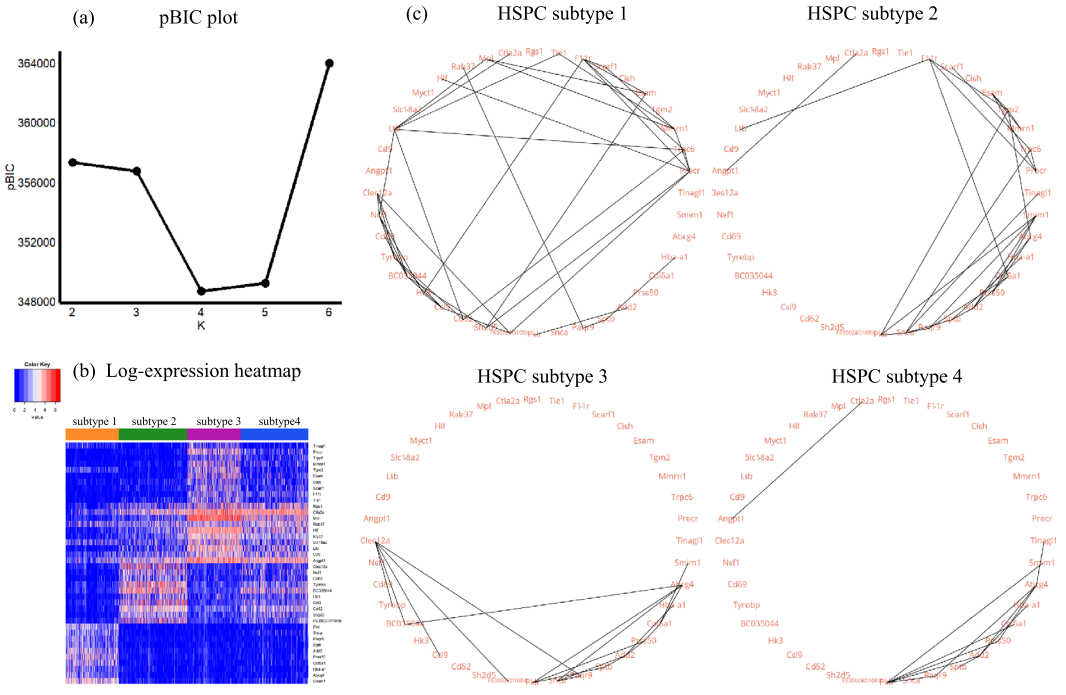

FIG. 4. *The results of real application* 1: (*a*) *The pBIC plot for K from* 2 *to* 6. (*b*) *The log-expression heatmap of the four estimated cell clusters on* 40 *marker genes.* (*c*) *Gene regulatory networks for the four HSPC subtypes.*

MCMC convergence diagnostic plots in Figure S11 of the Supplementary Material (Wu and Luo (2022)) show that the chain has attained stationary, and the four estimated precision matrices are shown in Figure S12 of the Supplementary Material (Wu and Luo (2022)). Figure 4(b) is the log-expression heatmap of the four estimated cell clusters on 40 marker genes, showing clear differential expression patterns. Specifically, following the cell annotations in Nestorowa et al. (2016), HSPC cell subtype 1 is mostly composed of megakaryocyte-erythrocyte progenitors (MEP), and subtype 3 mainly consists of long-term hematopoietic stem cells (LT-HSC). Subtypes 2 and 4 both represent a mix of other types of progenitor cells.

Figure 4(c) displays gene regulatory networks for the four HSPC subtypes. We observe that gene-gene connections vary across the four subtypes and the network in subtype 1 is more dense than in other three subtypes. For example, the edge between *Mpl* and *Esam* is present in subtype 1 (MEP) but absent in other subtypes, indicating that *Mpl* and *Esam* may have direct effects given other 38 genes. A previous study (Kohlscheen et al. (2015)) identified *Esam* as one of the downstream effectors of Thpo/Mpl-signaling in HSC, so our finding provides the evidence that the regulation effect may be also in MEP cells. In addition, we identified a *Snca—Add2* link in subtype 3 (LT-HSC). Gajović et al. (2006) found that *Snca* is expressed in mouse embryonic stem cells with mutated *Add2* but is not in control cells with the null mutation of *Add2*, indicating that this phenomenon may also happen in mouse hematopoietic stem cells.

The tail probabilities of the differential partial correlation compared to HSPC subtype 1 are also reported in Figure S13 of the Supplementary Material (Wu and Luo (2022)). Using threshold 0.5, no edge with different signs of partial correlations is discovered. Finally, we carried out the model checking to test whether the proposed model can fit the real data well. A predictive sample $\mathbf{X}^{\text{pred}}$ was first simulated from the posterior predictive distribution $p(\mathbf{X}^{\text{pred}}|\mathbf{X})$ (Gelman et al. (2013)), where $\mathbf{X}^{\text{pred}}$ has the same shape as $\mathbf{X}$. Subsequently, we calculated the zero proportion for each cell in $\mathbf{X}^{\text{pred}}$ and $\mathbf{X}$ and then compared them in the histogram of cell-specific zero proportions (Figure S14 of the Supplementary Material (Wu and Luo (2022))), showing that the fitted model can produce a dataset with similar zero proprotion distributions to the observed dataset. Moreover, for each gene, we compared its predicted expressions to the observed expressions across cells. Figures S15–S18 of the Supplementary Material (Wu and Luo (2022)) indicate that the predicted samples have relatively large overlaps with the observations in most genes, so the proposed model has a satisfactory fit to the genewise marginal expression distributions.

5.2. *Human retina cell data*. We also applied our model to transcriptomic data of human retina cells (Menon et al. (2019)). As this dataset provides cell-type labels, we selected cells from two main cell types, bipolar cells and macroglia, and data were then normalized using the same step described above. We focused on 34 marker genes provided by Menon et al. (2019) for the two cell types and removed cells with zero proportions larger than 75% in these marker genes, leading to 4697 cells. The IQR of zero proportions of cells is [44.1%, 70.6%] with the maximum value 73.5%. The pBIC plot in Figure 5(a) gives the optimal cell-type number $K = 4$, and this is validated in Menon et al. (2019) where three subtypes in macroglia are identified. When $K = 4$, the comparison between pBIC = 918,184.0 for tied precision matrices and pBIC = 529,741.8 for distinct precision matrices shows the existence of heterogeneity among precision matrices.

After the application of our model, the trace plots in Figure S19 of the Supplementary Material (Wu and Luo (2022)) show the MCMC chain has converged, and the four precision matrices are also displayed in Figure S20 of the Supplementary Material (Wu and Luo (2022)). Figure 5(b) is the the heatmap of log-expression values where cell types were annotated using names in Menon et al. (2019). The ARI of two major cell types between clustering

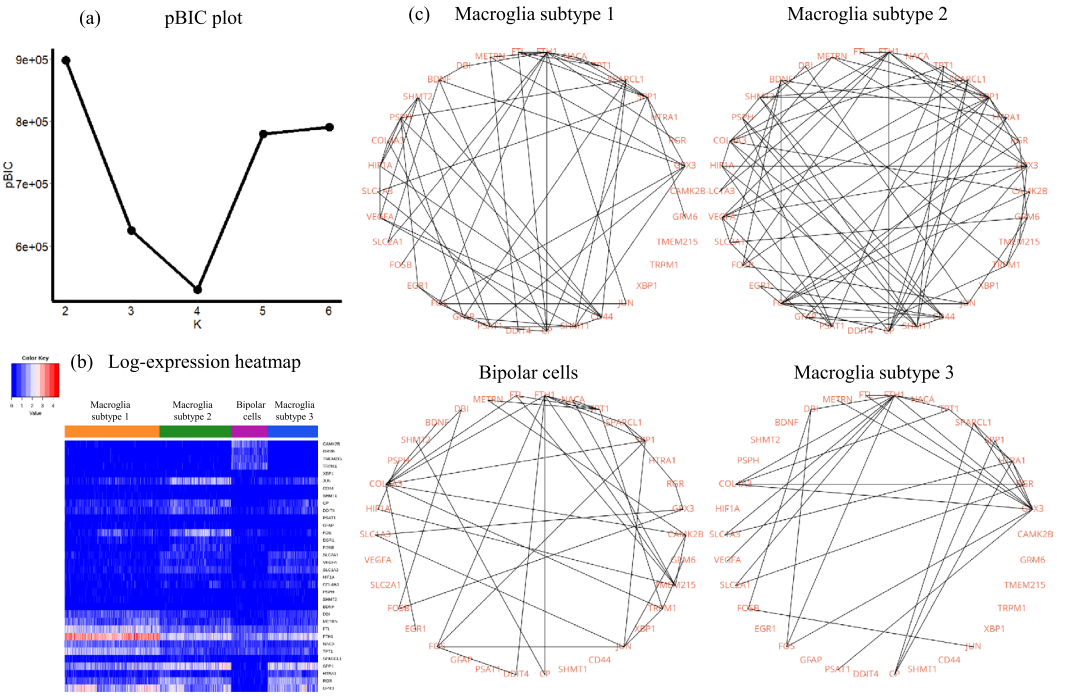

FIG. 5. *The results of real application* 2: (*a*) *The pBIC plot for K from* 2 *to* 6. (*b*) *The log-expression heatmap of the four cell clusters on* 34 *marker genes*. (*c*) *Gene regulatory networks for the four subtypes*.

result and the ACTIONet estimation (Mohammadi et al. (2018)) used in Menon et al. (2019) is 0.99, indicating that the two methods have similar performance in terms of clustering. Gene regulatory networks were shown in Figure 5(c). We can see that the gene regulatory networks in Macroglia subtypes 1 and 2 are more dense than those in bipolar cells and macroglia subtype 3. One finding is that the *GRM6—TRPM1* edge is present in only bipolar cells, and a genetic study (van Genderen et al. (2009)) confirms it by claiming that *TRPM1* is controlled by the *GRM6* signaling cascade in retina bipolar cells.

Moreover, the network structures can provide more insights into differentially expressed genes. For example, we can observe from the expression heatmap in Figure 5(b) that *FOS* is a marker for macroglia subtype 2, while it attains the maximal degree 9 in subtype 2 (5 in subtype 1 and 4 in subtype 3). Similarly, *FTH1* is a marker for macroglia subtype 1 in terms of expression, and it also has the mamximal degree 12 in that subtype (9 in subtype 2 and 8 in subtype 3). Thus, marker genes found by differential expressions may also exhibit differences in the network property.

Finally, using the bipolar cells as the reference, we identified the following edges with differential partial correlations (Figure S21 of the Supplementary Material (Wu and Luo (2022))): *FTH1—CP*, *FTH1—FOS* and *FTH1—SPP1* in macroglia subtype 1, *FTH1—CP* in macroglia subtype 2 and *DBI—FOSB* in macroglia subtype 3. We found that all the reported edges have negative partial correlations in bipolar cells, while they have positive signs in corresponding macroglia subtypes, indicating that the gene-gene correlation signs may also contribute to the cell heterogeneity. Finally, following the similar model-checking procedure above, Figures S22–S26 of the Supplementary Material (Wu and Luo (2022)) support the good fit of the proposed model to the observed data.

6. Discussion. We presented a Bayesian approach to simultaneously discover cell types and estimate cell-type-specific gene regulatory network for zero-inflated single-cell expres-

sion data. The graphical spike-slab prior was employed to induce sparsity of the gene regulatory networks. An efficient MCMC sampling scheme was developed to conduct posterior inference. The model outperforms competing statistical graphical models and is also robust to model-misspecified case via simulation study.

In the implementation of BLGGM, we assume that the scRNA-seq raw count data are first normalized to account for library sizes (e.g., counts per the median library size of cells), resulting in the input data matrix $\mathbf{X}$ whose elements are continuous and nonnegative values. Since the normalization procedure has considered the library size/sequencing depth issue, we do not need to take the library size into account when modeling $\mathbf{X}$.

There are several directions to improve the current work. For example, it is possible to directly model scRNA-seq raw counts using count-based distributions, such as zero-inflated negative-binomial/Poisson-log-normal distribution. We currently used continuous log-normal distribution for the following reasons. First, the normalization for read counts is usually a standard step in scRNA-seq data analysis pipeline, such as Seurat. However, normalized scRNA-seq data are not count-based anymore, and the empirical distribution exhibits asymmetry. The two features can be well captured by the log-normal distribution which is both continuous and asymmetric. Second, using the continuous distribution is more computationally efficient than the discrete Poisson or negative binomial distribution in Bayesian posterior sampling. Third, as discussed in the model misspecification part, by fitting transformed data via log-normal, the performance is still satisfactory. Fourth, in scRNA-seq data analysis literature, some well-known approaches are also based on continuous distributions, such as zero-inflated normal (Pierson and Yau (2015)), gamma (Lin et al. (2020)) and normal (Chen and Zhou (2017)).

In the proposed model the dropout probability for gene g in cell i is $\Phi(\lambda_{g0} + \lambda_{g1}\theta_{gi})$, and the quantity can explain the zero proportions in real data to some extent based on the posterior model checking. To make the zero inflation explained by the model more realistic and dynamic, it is straightforward to design an extension of BLGGM by assuming that the dropout coefficients λ_{g0} and λ_{g1} depend on the cell-type label k. Subsequently, for cells in cell type k, their zero proportion for gene g becomes $\Phi(\lambda_{g0,k} + \lambda_{g1,k}\theta_{gi})$, so the number of zeros can be more dynamic across cell types. We leave it for our future work.

As discussed in Section 2, BLGGM aims to recover the local conditional independence for selected genes. Sometimes, the local conditional independence can be equivalent to the global independence if the selected genes are independent of the filtered genes (i.e., $\mathbf{\Omega}^*_{12} = \mathbf{\Omega}^*_{21} = 0$). In this case, $\mathbf{\Omega}_{p\times p} = \mathbf{\Omega}^*_1$. In general, there have been some works (Choi et al. (2011), Meng, Eriksson and Hero (2014)) to recover $\mathbf{\Omega}^*_1$, based on $\mathbf{\Omega}_{p\times p}$ through optimization strategies, so the precision matrix estimates given by BLGGM can be used as the input for them to infer the global conditional independence.

To improve the computation efficiency of BLGGM, the EM algorithm can be applied to estimate the precision matrices. However, the EM algorithm needs to be implemented in a variational way (Bishop (2006)) because in the E step the conditional density of the latent variables $\boldsymbol{\theta}$, given observations $\mathbf{X}$, $p(\boldsymbol{\theta}|\mathbf{X}, -)$, does not have an analytical form. Considering that, in practice, the EM algorithm is more efficient than the MCMC sampling as EM only searches the modes rather than capture the whole distribution and the variational EM can further boost the speed, we expect that the EM implementation of BLGGM can scale to a much larger number of genes.

Acknowledgments. We are grateful to the Editor and three reviewers for their constructive and invaluable comments which have greatly improved the quality of the paper. We thank the High-performance Computing Platform of Renmin University of China for providing computing resources.

Funding. Xiangyu Luo was supported in part by National Natural Science Foundation of China (11901572), the Fundamental Research Funds for the Central Universities, the Research Funds of Renmin University of China (19XNLG08) and the fund for building world-class universities (disciplines) of Renmin University of China.

SUPPLEMENTARY MATERIAL

Additional details, analyses, and results (DOI: 10.1214/21-AOAS1582SUPPA; .pdf). This file contains supplementary sections, tables, and figures that provide additional details, analyses, and results.

Code and data (DOI: 10.1214/21-AOAS1582SUPPB; .zip). This file contains R code and datasets to reproduce results in simulation and real application.

REFERENCES

AIBAR, S., GONZÁLEZ-BLAS, C. B., MOERMAN, T., IMRICHOVA, H., HULSELMANS, G., RAMBOW, F., MARINE, J.-C., GEURTS, P., AERTS, J. et al. (2017). SCENIC: Single-cell regulatory network inference and clustering. *Nat. Methods* **14** 1083–1086.

AMEMIYA, T. (1984). Tobit models: A survey. *J. Econometrics* **24** 3–61. MR0739428 https://doi.org/10.1016/0304-4076(84)90074-5

BISHOP, C. M. (2006). *Pattern Recognition and Machine Learning*. *Information Science and Statistics*. Springer, New York. MR2247587 https://doi.org/10.1007/978-0-387-45528-0

CHATTERJEE, S., KAPOOR, A., AKIYAMA, J. A., AUER, D. R., LEE, D., GABRIEL, S., BERRIOS, C., PENNACCHIO, L. A. and CHAKRAVARTI, A. (2016). Enhancer variants synergistically drive dysfunction of a gene regulatory network in Hirschsprung disease. *Cell* **167** 355–368.

CHEN, M. and ZHOU, X. (2017). Controlling for confounding effects in single cell RNA sequencing studies using both control and target genes. *Sci. Rep.* **7** 1–14.

CHOI, M. J., TAN, V. Y. F., ANANDKUMAR, A. and WILLSKY, A. S. (2011). Learning latent tree graphical models. *J. Mach. Learn. Res.* **12** 1771–1812. MR2813153

CSARDI, G., NEPUSZ, T. et al. (2006). The igraph software package for complex network research. *InterJournal*, *Complex Syst.* **1695** 1–9.

DANAHER, P., WANG, P. and WITTEN, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 373–397. MR3164871 https://doi.org/10.1111/rssb.12033

DEMPSTER, A. P. (1972). Covariance selection. *Biometrics* **28** 157–175. MR3931974

DOBRA, A., LENKOSKI, A. and RODRIGUEZ, A. (2011). Bayesian inference for general Gaussian graphical models with application to multivariate lattice data. *J. Amer. Statist. Assoc.* **106** 1418–1433. MR2896846 https://doi.org/10.1198/jasa.2011.tm10465

EDGAR, R., DOMRACHEV, M. and LASH, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30** 207–210.

FRALEY, C. and RAFTERY, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Statist. Assoc.* **97** 611–631. MR1951635 https://doi.org/10.1198/016214502760047131

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

GAJOVIĆ, S., MITREČIĆ, D., AUGUSTINČIĆ, L., IACONCIG, A. and MURO, A. F. (2006). Unexpected rescue of alpha-synuclein and multimerin1 deletion in C57BL/6JOlaHsd mice by beta-adducin knockout. *Transgenic Res.* **15** 255–259.

GALLOPIN, M., RAU, A. and JAFFRÉZIC, F. (2013). A hierarchical Poisson log-normal model for network inference from RNA sequencing data. *PLoS ONE* **8** e77503. https://doi.org/10.1371/journal.pone.0077503

GAN, L., YANG, X., NARISETTY, N. and LIANG, F. (2019). Bayesian joint estimation of multiple graphical models. In *Advances in Neural Information Processing Systems* 9799–9809.

GAO, C., ZHU, Y., SHEN, X. and PAN, W. (2016). Estimation of multiple networks in Gaussian mixture models. *Electron. J. Stat.* **10** 1133–1154. MR3499523 https://doi.org/10.1214/16-EJS1135

GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2013). *Bayesian Data Analysis*, 3rd ed. *Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. MR3235677

GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **6** 721–741.

GUO, J., LEVINA, E., MICHAILIDIS, G. and ZHU, J. (2011). Joint estimation of multiple graphical models. *Biometrika* **98** 1–15. MR2804206 https://doi.org/10.1093/biomet/asq060

HAO, B., SUN, W. W., LIU, Y. and CHENG, G. (2018). Simultaneous clustering and estimation of heterogeneous graphical models. *J. Mach. Learn. Res.* **18** Paper No. 217, 58 pp. MR3827105

HUBERT, L. and ARABIE, P. (1985). Comparing partitions. *J. Classification* **2** 193–218.

KIM, S. (2015). ppcor: An R package for a fast calculation to semi-partial correlation coefficients. *Commun. Stat. Appl. Methods* **22** 665.

KOHLSCHEEN, S., WINTTERLE, S., SCHWARZER, A., KAMP, C., BRUGMAN, M. H., BREUER, D. C., BÜSCHE, G., BAUM, C. and MODLICH, U. (2015). Inhibition of Thrombopoietin/Mpl signaling in adult hematopoiesis identifies new candidates for hematopoietic stem cell maintenance. *PLoS ONE* **10** e0131866. https://doi.org/10.1371/journal.pone.0131866

LI, Z., MCCORMICK, T. and CLARK, S. (2019). Bayesian joint spike-and-slab graphical lasso. In *International Conference on Machine Learning* 3877–3885.

LIN, Z., WANG, T., YANG, C. and ZHAO, H. (2017). On joint estimation of Gaussian graphical models for spatial and temporal data. *Biometrics* **73** 769–779. MR3713111 https://doi.org/10.1111/biom.12650

LIN, Z., ZAMANIGHOMI, M., DALEY, T., MA, S. and WONG, W. H. (2020). Model-based approach to the joint analysis of single-cell data on chromatin accessibility and gene expression. *Statist. Sci.* **35** 2–13. MR4071354 https://doi.org/10.1214/19-STS714

LUO, X. and WEI, Y. (2018). Nonparametric Bayesian learning of heterogeneous dynamic transcription factor networks. *Ann. Appl. Stat.* **12** 1749–1772. MR3852696 https://doi.org/10.1214/17-AOAS1129

MA, J. and MICHAILIDIS, G. (2016). Joint structural estimation of multiple graphical models. *J. Mach. Learn. Res.* **17** Paper No. 166, 48 pp. MR3555057

MCDAVID, A., GOTTARDO, R., SIMON, N. and DRTON, M. (2019). Graphical models for zero-inflated single cell gene expression. *Ann. Appl. Stat.* **13** 848–873. MR3963555 https://doi.org/10.1214/18-AOAS1213

MENG, Z., ERIKSSON, B. and HERO, A. (2014). Learning latent variable Gaussian graphical models. In *International Conference on Machine Learning* 1269–1277. PMLR.

MENON, M., MOHAMMADI, S., DAVILA-VELDERRAIN, J., GOODS, B. A., CADWELL, T. D., XING, Y., STEMMER-RACHAMIMOV, A., SHALEK, A. K., LOVE, J. C. et al. (2019). Single-cell transcriptomic atlas of the human retina identifies cell types associated with age-related macular degeneration. *Nat. Commun.* **10** 1–9.

METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. and TELLER, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21** 1087–1092.

MIAO, W., DING, P. and GENG, Z. (2016). Identifiability of normal and normal mixture models with nonignorable missing data. *J. Amer. Statist. Assoc.* **111** 1673–1683. MR3601726 https://doi.org/10.1080/01621459.2015.1105808

MOHAMMADI, A. and WIT, E. C. (2015). Bayesian structure learning in sparse Gaussian graphical models. *Bayesian Anal.* **10** 109–138. MR3420899 https://doi.org/10.1214/14-BA889

MOHAMMADI, R. and WIT, E. C. (2019). BDgraph: An R package for Bayesian structure learning in graphical models. *J. Stat. Softw.* **89** 1–30.

MOHAMMADI, S., RAVINDRA, V., GLEICH, D. F. and GRAMA, A. (2018). A geometric approach to characterize the functional identity of single cells. *Nat. Commun.* **9** 1–10.

MORITA, Y., EMA, H. and NAKAUCHI, H. (2010). Heterogeneity and hierarchy within the most primitive hematopoietic stem cell compartment. *J. Exp. Med.* **207** 1173–1182.

NEAL, R. M. (2011). MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo. Chapman & Hall/CRC Handb. Mod. Stat. Methods* 113–162. CRC Press, Boca Raton, FL. MR2858447

NESTOROWA, S., HAMEY, F. K., PIJUAN SALA, B., DIAMANTI, E., SHEPHERD, M., LAURENTI, E., WILSON, N. K., KENT, D. G. and GÖTTGENS, B. (2016). A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood, J. Amer. Soc. Hematol.* **128** e20–e31.

NEWTON, M. A., NOUEIRY, A., SARKAR, D. and AHLQUIST, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5** 155–176.

NTRANOS, V., YI, L., MELSTED, P. and PACHTER, L. (2019). A discriminative learning approach to differential expression analysis for single-cell RNA-seq. *Nat. Methods* **16** 163–166. https://doi.org/10.1038/s41592-018-0303-9

PAN, W. and SHEN, X. (2007). Penalized model-based clustering with application to variable selection. *J. Mach. Learn. Res.* **8** 1145–1164.

PETERSON, C., STINGO, F. C. and VANNUCCI, M. (2015). Bayesian inference of multiple Gaussian graphical models. *J. Amer. Statist. Assoc.* **110** 159–174. MR3338494 https://doi.org/10.1080/01621459.2014.896806

PIERSON, E. and YAU, C. (2015). ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.* **16** 1–10.

PRATAPA, A., JALIHAL, A. P., LAW, J. N., BHARADWAJ, A. and MURALI, T. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat. Methods* **17** 147–154.

QIU, X., RAHIMZAMANI, A., WANG, L., MAO, Q., DURHAM, T., MCFALINE-FIGUEROA, J. L., SAUNDERS, L., TRAPNELL, C. and KANNAN, S. (2018). Towards inferring causal gene regulatory networks from single cell expression measurements. *BioRxiv* 426981.

REN, M., ZHANG, S., ZHANG, Q. and MA, S. (2021a). Gaussian graphical model-based heterogeneity analysis via penalized fusion. *Biometrics*. https://doi.org/10.1111/biom.13426

REN, M., ZHANG, S., ZHANG, Q. and MA, S. (2021b). HeteroGGM: An R package for Gaussian graphical model-based heterogeneity analysis. *Bioinformatics* **37** 3073–3074.

RISSO, D., PERRAUDEAU, F., GRIBKOVA, S., DUDOIT, S. and VERT, J.-P. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* **9** 1–17.

RODRÍGUEZ, A., LENKOSKI, A. and DOBRA, A. (2011). Sparse covariance estimation in heterogeneous samples. *Electron. J. Stat.* **5** 981–1014. MR2836767 https://doi.org/10.1214/11-EJS634

ROZENBLATT-ROSEN, O., STUBBINGTON, M. J., REGEV, A. and TEICHMANN, S. A. (2017). The Human Cell Atlas: From vision to reality. *Nat. News* **550** 451.

SAEGUSA, T. and SHOJAIE, A. (2016). Joint estimation of precision matrices in heterogeneous populations. *Electron. J. Stat.* **10** 1341–1392. MR3507368 https://doi.org/10.1214/16-EJS1137

SONG, F., CHAN, G. M. A. and WEI, Y. (2020). Flexible experimental designs for valid single-cell RNA-sequencing experiments allowing batch effects correction. *Nat. Commun.* **11** 1–15.

TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* **82** 528–550. MR0898357

VAN GENDEREN, M. M., BIJVELD, M. M., CLAASSEN, Y. B., FLORIJN, R. J., PEARRING, J. N., MEIRE, F. M., MCCALL, M. A., RIEMSLAG, F. C., GREGG, R. G. et al. (2009). Mutations in TRPM1 are a common cause of complete congenital stationary night blindness. *Am. J. Hum. Genet.* **85** 730–736.

VIETH, B., PAREKH, S., ZIEGENHAIN, C., ENARD, W. and HELLMANN, I. (2019). A systematic evaluation of single cell RNA-seq analysis pipelines. *Nat. Commun.* **10** 1–11.

WANG, H. (2012). Bayesian graphical lasso models and efficient posterior computation. *Bayesian Anal.* **7** 867–886. MR3000017 https://doi.org/10.1214/12-BA729

WANG, H. (2015). Scaling it up: Stochastic search structure learning in graphical models. *Bayesian Anal.* **10** 351–377. MR3420886 https://doi.org/10.1214/14-BA916

WANG, H. and LI, S. Z. (2012). Efficient Gaussian graphical model determination under G-Wishart prior distributions. *Electron. J. Stat.* **6** 168–198. MR2879676 https://doi.org/10.1214/12-EJS669

WU, Q. and LUO, X. (2022). Supplement to "Estimating heterogeneous gene regulatory networks from zero-inflated single-cell expression data." https://doi.org/10.1214/21-AOAS1582SUPPA, https://doi.org/10.1214/21-AOAS1582SUPPB

YANG, H.-J., RATNAPRIYA, R., COGLIATI, T., KIM, J.-W. and SWAROOP, A. (2015). Vision from next generation sequencing: Multi-dimensional genome-wide analysis for producing gene regulatory networks underlying retinal development, aging and disease. *Prog. Retin. Eye Res.* **46** 1–30.

YUAN, M. and LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. MR2367824 https://doi.org/10.1093/biomet/asm018

ZHANG, H., XU, J., JIANG, N., HU, X. and LUO, Z. (2015). PLNseq: A multivariate Poisson lognormal distribution for high-throughput matched RNA-sequencing read count data. *Stat. Med.* **34** 1577–1589. MR3334677 https://doi.org/10.1002/sim.6449