Journal of the Royal Statistical Society Series C: Applied Statistics, 2025, **74**, 204–228
https://doi.org/10.1093/jrsssc/qlae052
Advance access publication 14 October 2024
Original Article

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series C
Applied Statistics

Sparse inference of the human haematopoietic system from heterogeneous and partially observed genomic data

Gianluca Sottile[1], Luigi Augugliaro[1], Veronica Vinciotti[2], Walter Arancio[3,4] and Claudia Coronnello[4]

[1]Department of Economics, Business and Statistics, University of Palermo, Palermo, Italy
[2]Department of Mathematics, University of Trento, Trento, Italy
[3]Institute for Biomedical Research and Innovation, National Research Council, Palermo, Italy
[4]Advanced Data Analysis Group, Fondazione Ri.MED, Palermo, Italy

Address for correspondence: Gianluca Sottile, Department of Economics, Business and Statistics, University of Palermo, Viale delle Scienze, 8, 90128 Palermo, Italy. Email: gianluca.sottile@unipa.it

Abstract

Haematopoiesis is the process of blood cells' formation, with progenitor stem cells differentiating into mature forms such as white and red blood cells or platelets. While progenitor cells share regulatory pathways involving common nuclear factors, specific networks shape their fate towards particular lineages. This paper analyses the complex regulatory network that drives the formation of mature red blood cells and platelets from their common precursors. Using the latest reverse transcription quantitative real-time PCR genomic data, we develop a dedicated graphical model that incorporates the effect of external genomic data and allows inference of regulatory networks from the high-dimensional and partially observed data.

Keywords: graphical lasso, heterogeneous data, high-dimensional data, human haematopoietic system, multiple Gaussian graphical models, missing data

1 Introduction

Human haematopoiesis is a complex cellular process that occurs continuously during the human lifetime and that produces cellular blood components from haematopoietic stem cells. In dedicated niches in the bone marrow, a mixed population of precursors at different levels of differentiation coexists with a functional stroma. Via cross-regulation, it physiologically orchestrates the maintenance of stem-like immature precursors and the commitment towards specific cell lineages, together with the differentiation, the maturation and the release of the mature forms. Recent studies have focused on the megakaryocyte-erythroid progenitor (MEP) (Psaila et al., 2016). The MEP starts from an early immature precursor (pre-MEP) and then commits itself towards either an erythroid lineage (E-MEP), that will become red blood cells, or a megakaryocyte lineage (MK-MEP), that will produce mature platelets. While these precursors have many differentiation pathways in common, which are driven by key proteins called nuclear factors, specific networks of regulation shape their fate towards one lineage or another. The precise disentanglement of the pathways governing the commitment of the two populations is a difficult task and has received recent attention (Doré & Crispino, 2011).

While nuclear factors are key players in the late stages of haematopoiesis, they do not act in isolation. Many studies have shown how membrane-bound receptors are also important in the identification and classification of haematopoietic cells (Cheng et al., 2020). Indeed, membrane-bound receptors play a pivotal role in the relationship with the surrounding cellular microenvironment,

Received: June 6, 2022. **Revised:** April 23, 2024. **Accepted:** September 16, 2024

and their activation is transduced to the nucleus to reprogram or fine tune the cell fate. As well as inferring the network of regulation among nuclear factors, the proposed method aims to quantify the effect of these membrane-bound receptors on the nuclear factor activities. From a modelling point of view, this creates a rather complex picture with a number of challenging features: firstly, data at the level of nuclear activities (response variables) and at the level of membrane receptor activities (predictor variables or generally external covariates), with potentially high dimensionality on both; secondly, dependence structures both between and within the two sets of variables; thirdly, the presence of multiple biological conditions (Pre-MEP, E-MEP, and MK-MEP) which may have many features in common but also some key differences.

An added layer of complexity comes from the fact that the reverse transcription quantitative real-time PCR (RT-qPCR) data that will be used for the study, generated by Psaila et al. (2016), is characterized by a large percentage of missing data, which is typical for this type of data. In the literature, no consensus exists on how to model the missing process for the unobserved RT-qPCR data. The most common explanation is that some samples fail to attain the minimum (prespecified) signal intensity before reaching the maximum number of cycles (denoted with Ct for threshold-cycle and with a maximum of 40 for the case of Psaila et al. (2016)). In this case, the missing data are generated by a right-censoring mechanism, as their true Ct value is greater than the maximum detection limit. This assumption underlies a number of methods for imputing missing data generated by RT-qPCR, starting from the simplest and most common approach of setting those values to the limit of detection, as in Psaila et al. (2016), to more sophisticated approaches that, for example, exploit the availability of technical or biological replicates (Boyer et al., 2013), account for the uncertainty in the missing data mechanism (Sherina et al., 2020) or allow for the simultaneous imputation of all transcripts from their joint multivariate Gaussian distribution (Augugliaro, Abbruzzo, et al., 2020). However, various authors have also pointed out how the censoring mechanism only partially explains the extent of missingness typically observed in these data (McCall et al., 2014). This suggests that some missing data may result from a failed amplification and, therefore, their true value could also be lower than the detection limit. As we shall discuss in Section 2, we find evidence of this in the data of Psaila et al. (2016).

Motivated by the applied setting of blood cell formation described above, in this paper we develop a graphical modelling approach, with potentially missing or censored data, that accounts for lineage-dependent regulatory networks, both at the level of the response variables (nuclear activities) and of the covariates (membrane receptor activities) as well as lineage-dependent associations between the two. To estimate the proposed model under a high-dimensional regime, we develop an extension of the penalized estimators developed for the standard and conditional Gaussian graphical model setting under missingness (Augugliaro, Abbruzzo, et al., 2020; Augugliaro, Sottile, et al., 2020; Augugliaro et al., 2023; Städler & Bühlmann, 2012) to the case of multiple conditional Gaussian graphical models. Imposing sparsity both within and between the response and predictor levels provides a challenging computational setting: we propose a careful combination of regularized inferential procedures at the different levels, which we solve based on a suitably developed and efficient alternating direction method of multipliers (ADMM) algorithms (Boyd et al., 2011). On one side, the methodology aims to identify a common shared network among all the differentiation states. On the other side, it aims to elucidate specific features that characterize the most undifferentiated state (Pre-MEP) and each of the committed lineages (E-MEP and MK-MEP).

1.1 Related works

Inference of the dependence structure of random variables, such as the nuclear activities in the applied setting just described, falls within the remit of graphical models. Among these, Gaussian graphical models (GGMs) are commonly used for multivariate continuous random variables, thanks to their ability to depict highly complex conditional dependence structures among the random variables in a parsimonious way. Among the penalized estimators proposed in the literature, the graphical lasso (glasso) (Yuan & Lin, 2007) is undoubtedly the most famous one. Recently, the usage of complex sampling schemes and experimental settings have motivated various extensions to account for the heterogeneity in the observed data.

A first strand of extensions has looked at the case of data measured from different sources or under different conditions, such as the undifferentiated cellular state and the two committed lineages in the

formation of blood cells described above. In these cases, a comprehensive description of the data-generating process can be obtained by fitting a collection of GGMs, otherwise called a *multiple GGM*, with one network associated to each condition but with the different networks sharing some common structures, such as the presence or intensity of the dependences. To accomplish this joint estimation problem, a number of authors have developed specific penalized approaches (Danaher et al., 2014; Guo et al., 2011; Kling et al., 2015; Lee & Liu, 2015; Lin et al., 2016; Ma & Michailidis, 2016; Majumdar & Michailidis, 2022; Samanta et al., 2022; Xie et al., 2016; Zhang et al., 2017; Zhu et al., 2014). Of particular notice for the method proposed in this paper is the double-penalized estimator, which we refer to as joint glasso, introduced by Danaher et al. (2014).

The joint glasso approach focuses mainly on network estimation and treats the expected values as nuisance parameters. A second strand of extensions has concentrated on modelling the expected values, by capturing dependences with external data at the mean level. Penalized methods developed to analyse this kind of heterogeneous data are built on the notion of *conditional GMMs* (Lafferty et al., 2001), which, in turn, are based on the assumption that the covariates affect the multivariate density only through a linear model on the expected value. The penalized estimators that have been proposed to explore the sparse structure of a conditional GGM typically use two specific penalty functions: one that induces sparsity in the regression coefficient matrix, capturing the association between the response variables and the external covariates, and a second one that encourages sparsity in the dependence structure between the response variables (Chen et al., 2016; Chiquet et al., 2017; Li et al., 2012; Rothman et al., 2010; Wang, 2015; Yin & Li, 2011, 2013). These methods normally assume that the network describing the dependences between the response variables is independent of the covariates. A small number of studies have proposed further extensions of these estimators to the case of *multiple conditional GGMs*, essentially providing a bridge with the multiple GGM methods described above (Chun et al., 2013; Huang et al., 2018; Majumdar & Michailidis, 2022). In all these methods, the covariates are not treated as random variables and their dependence structure is not explored.

1.2 Overview of the paper

In Section 2, we present the motivating example and provide some summary statistics. Section 3 describes the details of the proposed Gaussian graphical model and derives the likelihood under the case of partially observed data, both at the response and covariate levels. Section 4 presents the penalized likelihood that is used as objective function. Section 5 details the steps of the EM algorithm and of the ADMM algorithms that make up the different sub-routines for optimization of the penalized likelihood. Section 6 addresses the problem of how to select the tuning parameters. Section 7 describes an extensive simulation study where we evaluate the performance of the method in terms of network recovery and parameter estimation, and compare it with existing methods. In Section 8, we provide an application of the methodology to the inference of the haematopoietic system and, finally, Section 9 concludes with a final discussion. Online supplementary material provide the pseudo-code and additional details of the proposed algorithms.

2 Data description

In this paper, we focus on the process of blood cell formation described in the Introduction. We use the data generated by Psaila et al. (2016), as well as their classification of the $n = 681$ megakaryocyte-erythroid progenitor (MEP) cells into three distinct sub-populations, identified via a principle component analysis. A close inspection of the expression patterns of the megakaryocyte-associated and erythroid-associated genes across the three sub-populations have led Psaila et al. (2016) to the identification of the three groups as, respectively, the most undifferentiated state, which they name as 'precursor-MEP' (Pre-MEP, $n_1 = 255$ samples), and two committed lineages, which they name as 'erythroid-primed MEP' (E-MEP, $n_2 = 241$) and 'megakaryocyte-primed' MEP (MK-MEP, $n_3 = 185$).

We analyse these data further, by elucidating common and specific features that characterize the most undifferentiated state and each of the two committed lineages at the level of cellular pathways among important genes. Of the 87 transcripts that are profiled by Psaila et al. (2016), 11 transcripts are not expressed in at least one of the three groups and hence they are removed from this analysis, while two transcripts (*B2M* and *GAPDH*) are housekeeping genes and are

used for data normalization. The remaining 74 genes are classified into two blocks: on one side the transcripts that code for membrane-bound receptors, their ligands, cytoplasmic signal transducers and metabolic enzymes (these will enter the model as $q = 40$ external covariates), while, on the other side the transcripts, such as the cyclin-dependent kinases, that code for nuclear factors and master regulators of nuclear and cellular activities (resulting in a total of $p = 34$ response variables). When a protein possesses a dual activity (e.g. it can be both membrane-bound and nuclear) the preference is given to the nuclear activity. Using the method described in this paper, we aim to recover the lineage-dependent regulatory networks, both at the level of the response variables, of the covariates and of the associations between the two, and to identify the key lineage-specific features among the many shared pathways.

The expression of the transcripts in each cell is measured by RT-qPCR technology. As discussed in the Introduction, this technology can produce data characterized by a high percentage of missing i.e. some reactions fail to attain the minimum prespecified amount of signal. Among the various explanations for failing to attain a given Ct value, two are the most accredited ones. The first one is that, in a given experiment, the transcript fails to amplify, i.e. its amplification efficiency is poor. In this case, the true unobserved Ct value should be a value below the maximum detection limit. The second reason is that the true Ct value is a value above the upper detection limit and, consequently, it would be observed if the PCR was to run for more cycles. From a statistical modelling point-of-view, these two explanations correspond to two specific probabilistic models for the missing data mechanism, namely, the Missing-At-Random (MAR) in the first case and the censoring mechanism in the second case.

Although in Section 3, we will propose a general framework to handle both a censoring and an MAR mechanism, a choice needs to be made on any given dataset. For the data of Psaila et al. (2016) used in this study, we find that the censoring mechanism does not explain completely the extent of missingness, i.e. the percentage of nondetects far exceeds what one would expect from a censoring mechanism (McCall et al., 2014). Indeed, Figure 1. shows how, despite the percentage of missing data for each transcript tends to increase with the average cycle threshold value (black circles in the top panel), this percentage is far higher than the expected probability for a right censored mechanism and under a Gaussian assumption (red crosses in the top panel), calculated using the method described in Augugliaro et al. (2023). The bottom panel refers to one specific transcript and shows a typical example of how this discrepancy cannot be attributed solely to a potential miss-specification of the Gaussian distribution, given the significant gap that is observed between the measured threshold values and the limit of detection. Taking all of this into consideration, in Section 8, we perform the analysis by assuming a missing-at-random data mechanism, for both the responses and the covariates.

3 A Gaussian graphical model for partially observed heterogeneous data

As mentioned in the Introduction, for each sub-population, the model should capture the dependence structure within the nuclear factors, as well as that within the membrane-bound receptors and the associations between the two. The presence of missing data will require knowledge of the joint model for nuclear factors and membrane-bound receptors. This will motivate the selection of a specific form of parametrization, which will result in a combination of a standard Gaussian graphical model at the level of membrane receptor activities and a conditional Gaussian graphical model at the level of the nuclear activities conditional on the membrane receptor activities.

Consider the kth sub-population, which corresponds to one of the three biological conditions described in Section 2, i.e. Pre-MEP, E-MEP, and MK-MEP. We begin by formalizing the first GGM, i.e. the one related to the conditional dependence structure within the membrane-bound receptors. Let $X_{\cdot k} = (X_{1k}, \ldots, X_{qk})^\top$ be the q-dimensional random vector corresponding to the abundance of the q membrane receptor activities in sub-population k. Assume that it follows a multivariate Gaussian distribution, i.e. the density function takes the form:

$$\phi_x(x; \mu_k, \Omega_k) = (2\pi)^{-\frac{q}{2}} |\Omega_k|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}(x-\mu_k)^\top \Omega_k (x-\mu_k)\right\}, \tag{3.1}$$

where μ_k denotes the expected value of $X_{\cdot k}$ and $\Omega_k = \{\omega_{ijk}\}$, called precision matrix, denotes the inverse of the covariance matrix. In GGMs, the conditional dependence structure among the

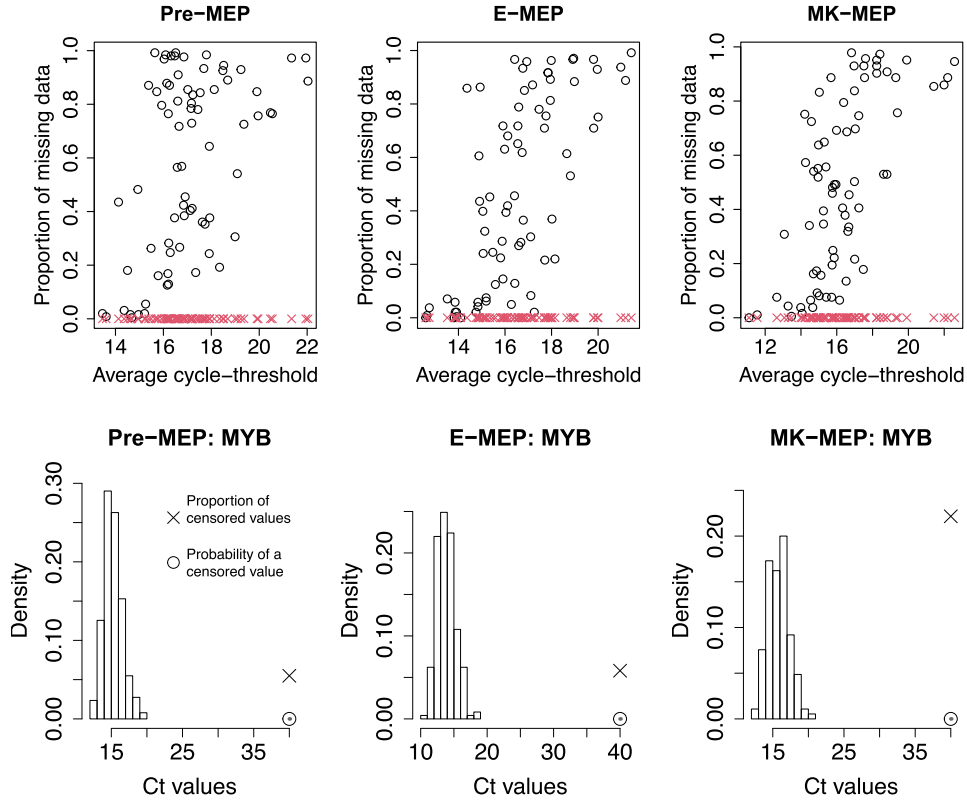

Figure 1. Top: For each transcript, the average Ct value is plotted against the proportion of missing data (circles) and the probability of observing censored values (crosses), calculated by the marginal distribution fitted using the method described in this paper. Bottom: Histograms of the observed Ct values for a selected nuclear factor *MYB* across the three sub-populations. Crosses indicate the proportion of censored values in the data, while circles indicate the expected probability of a censored value based on the fitted marginal model.

random variables is depicted by an undirected graph $G_{x,k} = \{V_x, E_{x,k}\}$, where V_x is the set $\{1, \ldots, q\}$ and $E_{x,k} \subseteq V_x \times V_x$ is the set of undirected edges. In GGMs, the random variables X_{ik} and X_{jk} are conditional independent given the remaining X_{lk} variables, with $l \neq i, j$, if the pair (i, j) is not an element of the edge set $E_{x,k}$. Using standard results on the factorization of the multivariate Gaussian distribution (see for example Lauritzen, 1996), it is possible to show that Ω_k and $G_{x,k}$ are intimately related to each other, as $(i, j) \neq E_{x,k}$ iff $\omega_{ijk} = 0$. This result plays a prominent role inside the theory of GGMs because it allows to turn the problem of inferring the conditional independence structure into the problem of estimating a sparse precision matrix.

The next step of the model specification process consists in specifying a conditional GGM that allows to infer how the membrane receptors affect the abundance of the nuclear factor activities and how these factors affect each other, after the effect of the membrane receptors is accounted for. To answer both questions, we assume that the conditional distribution of the p-dimensional random vector $Y_{\cdot k} = (Y_{1k}, \ldots, Y_{pk})^\top$, representing the abundance of the p nuclear factor activities in sub-population k, takes the form:

$$\phi_{y|x}(y \mid x; \beta_{0\cdot k}, B_k, \Theta_k) = (2\pi)^{-\frac{p}{2}} |\Theta_k|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}(y - \beta_{0\cdot k} - B_k^\top x)^\top \Theta_k (y - \beta_{0\cdot k} - B_k^\top x)\right\}, \qquad (3.2)$$

where $\beta_{0\cdot k}$ is the p-dimensional vector of intercepts, $B_k = (\beta_{1\cdot k} \cdots \beta_{p\cdot k})$ is the $q \times p$ regression coefficient matrix, and $\Theta_k = (\theta_{ijk})$ is the inverse of the conditional covariance matrix of $Y_{\cdot k}$ given $X_{\cdot k}$. In this conditional GGM, B_k represents the parametric tool for quantifying how the

membrane receptors affect the average abundance of the nuclear factor activities, whereas Θ_k allows to quantify how these factors affect each other, after the effect of the membrane receptors is accounted for. Indeed, like in a standard GGM, the factorization of the density (3.2) is related to the undirected graph $G_{y|x,k} = \{V_y, E_{y|x,k}\}$, where $V_y = \{1, \ldots, p\}$ and $E_{y|x,k} \subseteq V_y \times V_y$, whereby Y_{ik} and Y_{jk} are conditionally independent given $X_{\cdot k}$ and all the remaining Y_{lk} variables, with $l \neq i, j$, if the corresponding θ_{ijk} is zero.

The two GGMs presented above allow us to obtain valid inferential results on the conditional dependence structure among the variables of interest only when a fully observed dataset is available. As discussed in the Introduction, RT-qPCR data are typically only partially observed. In order to formalize a likelihood framework based on the notion of the observed likelihood function (see Little & Rubin, 2002 for more details), we use as starting point the joint distribution of the random vector $Z_{\cdot k} = (X_{\cdot k}^\top, Y_{\cdot k}^\top)^\top$, which is a multivariate Gaussian distribution whose expected value ψ_k, covariance matrix Σ_k and precision matrix Ξ_k can be partitioned as follows:

$$\psi_k = \begin{pmatrix} \psi_{x,k} \\ \psi_{y,k} \end{pmatrix}, \quad \Sigma_k = \begin{pmatrix} \Sigma_{xx,k} & \Sigma_{xy,k} \\ \Sigma_{yx,k} & \Sigma_{yy,k} \end{pmatrix}, \quad \text{and} \quad \Xi_k = \begin{pmatrix} \Xi_{xx,k} & \Xi_{xy,k} \\ \Xi_{yx,k} & \Xi_{yy,k} \end{pmatrix}. \tag{3.3}$$

The blocks of the matrix Ξ_k can be easily related to the parameters of the density (3.1) and (3.2). Therefore, this precision matrix contains all the necessary information needed to infer the graphs $G_{y|x,k}$ and $G_{x,k}$, respectively. First, note that by the identity $\Sigma_{xx,k}\Xi_{xy,k} + \Sigma_{xy,k}\Xi_{yy,k} = 0$ and the partitioned matrix inverse formula, the conditional expected value and conditional variance of $Y_{\cdot k}$ in (3.2) can be written as (Lauritzen, 1996)

$$\begin{aligned} \beta_{0\cdot k} + B_k^\top x &= \psi_{y,k} + \Sigma_{yx,k}\Sigma_{xx,k}^{-1}(x - \psi_{x,k}) = \psi_{y,k} - \Xi_{yy,k}^{-1}\Xi_{yx,k}(x - \psi_{x,k}), \\ \Theta_k^{-1} &= \Sigma_{yy,k} - \Sigma_{yx,k}\Sigma_{xx,k}^{-1}\Sigma_{xy,k} = \Xi_{yy,k}^{-1}. \end{aligned}$$

From this, it follows that $\Xi_{yy,k} = \Theta_k$ and $\Xi_{yx,k} = -\Theta_k B_k^\top$. Finally, recalling that $B_k = \Sigma_{xx,k}^{-1}\Sigma_{xy,k}$ and $\Xi_{yx,k} = -\Theta_k B_k^\top$, the identity $\Sigma_{xx,k}\Xi_{xx,k} + \Sigma_{xy,k}\Xi_{yx,k} = I$ implies that $\Xi_{xx,k} = \Omega_k + B_k \Theta B_k^\top$. Therefore, ψ_k and Ξ_k can be written as

$$\psi_k = \begin{pmatrix} \psi_{x,k} \\ \psi_{y,k} \end{pmatrix}, \quad \text{and} \quad \Xi_k = \begin{pmatrix} \Omega_k + B_k\Theta_k B_k^\top & -B_k\Theta_k \\ -\Theta_k B_k^\top & \Theta_k \end{pmatrix}. \tag{3.4}$$

In the remaining part of this paper, the density function of $Z_{\cdot k}$ will be denoted by $\phi_z(z; \mathrm{Y}_k)$, where $\mathrm{Y}_k = \{\psi_k, \Omega_k, B_k, \Theta_k\}$ is the full set of parameters of interest. The choice of parametrization (3.4), which is different to the one proposed in earlier related work (Augugliaro, Sottile, et al., 2020; Sohn & Kim, 2012), comes from the specific scientific questions of the current analysis, i.e. to account for, as well as distinguish, the dependence structures coming from the different sources of data. Finally, to include the probabilistic model for the missing data mechanism inside the proposed framework, by $R_{\cdot k}$ we denote the vector of the response indicator variables associated to $Z_{\cdot k}$, i.e. the $R_{jk} = 1$ iff Z_{jk} is missing.

Suppose that n observations are collected from $K \geq 2$ sub-populations and denote by $z_{i\cdot k}$ and $r_{i\cdot k}$ the ith realization of the random vectors $Z_{i\cdot k}$ and $R_{i\cdot k}$, which are assumed to be independent copies of $Z_{\cdot k}$ and $R_{\cdot k}$, respectively. Given $r_{i\cdot k}$, the vector $z_{i\cdot k}$ can be split into the observed and unobserved sub-vectors, which we denote with $z_{i\cdot k}^o$ and $z_{i\cdot k}^{no}$, respectively. The way the elements of $z_{i\cdot k}^{no}$ are treated depends closely on the probabilistic assumption used to model the missing data mechanism. As discussed above, we propose a general framework to handle both a censoring and an MAR mechanism, which we explain in the following.

First, recall that, under the assumption of independent sampling, the observed likelihood function is obtained by integrating $z_{i\cdot k}^{no}$ out of the joint distribution of $Z_{i\cdot k}$ and $R_{i\cdot k}$ (Little & Rubin, 2002). Formally,

$$L(\{\mathrm{Y}\}) = \prod_{k=1}^{K}\prod_{i=1}^{n_k} \int \phi(z_{i\cdot k}^o, z_{i\cdot k}^{no}; \mathrm{Y}_k)\Pr(R_{i\cdot k} = r_{i\cdot k} \mid z_{i\cdot k}^o, z_{i\cdot k}^{no})\mathrm{d}z_{i\cdot k}^{no}, \tag{3.5}$$

where $\{\mathrm{Y}\} = \{\mathrm{Y}_1, \ldots, \mathrm{Y}_K\}$ is the full set of parameters and n_k is the kth sample size. Suppose now that $z^{no}_{i\cdot k}$ comes from a censoring mechanism where $l_{\cdot k}$ and $u_{\cdot k}$ are the vectors of known lower and upper censoring values, respectively. Then, as discussed in Augugliaro, Abbruzzo, et al. (2020), the conditional probability of $R_{i\cdot k}$ given $z_{i\cdot k}$ is equal to

$$\Pr(R_{i\cdot k} = r_{i\cdot k} \mid z^{o}_{i\cdot k}, z^{no}_{i\cdot k}) = I(z^{o}_{i\cdot k} \in [l^{o}_{\cdot k}, u^{o}_{\cdot k}]) \times I(z^{no}_{i\cdot k} \in D_{i\cdot k}), \tag{3.6}$$

where $D_{i\cdot k}$, is defined as the Cartesian product of sub-regions D_{ijk}, whose definition depends on whether z^{no}_{ijk} is left or right censored. In particular, in the first case, i.e. when $z_{ijk} = l_{jk}$, the interval D_{ijk} is equal to $(-\infty, l_{jk})$ whereas, in the second case, i.e. when $z_{ijk} = u_{jk}$, we have that $D_{ijk} = (u_{jk}, +\infty)$. Putting (3.6) into (3.5), we have:

$$L(\{\mathrm{Y}\}) = \prod_{k=1}^{K} \prod_{i=1}^{n_k} \int_{D_{i\cdot k}} \phi(z^{o}_{i\cdot k}, z^{no}_{i\cdot k}; \mathrm{Y}_k) \mathrm{d}z^{no}_{i\cdot k}. \tag{3.7}$$

If instead $z^{no}_{i\cdot k}$ comes from an MAR mechanism, then, using the notation in Little (2021), we have that $\Pr(R_{i\cdot k} = r_{i\cdot k} \mid z^{o}_{i\cdot k}, z^{no}_{i\cdot k}) = \Pr(R_{i\cdot k} = r_{i\cdot k} \mid z^{o}_{i\cdot k}, \tilde{z}^{no}_{i\cdot k})$, for all $z^{no}_{i\cdot k}$ and $\tilde{z}^{no}_{i\cdot k}$, consequently the conditional probability of $R_{i\cdot k}$ can be taken out from the integral in (3.5). This implies that the resulting observed likelihood function is equal to (3.7) but with $D_{ijk} = \mathbb{R}$. In other words, a change between censoring and MAR mechanism affects expression (3.7) solely by how the intervals D_{ijk} are defined. Finally, it is worth noting that the definition (3.7) is valid also under the assumption of Missing-Completely-At-Random (MCAR). Indeed, this assumption implies that $\Pr(R_{i\cdot k} = r_{i\cdot k} \mid z_{i\cdot k}) = \Pr(R_{i\cdot k} = r_{i\cdot k} \mid \tilde{z}_{i\cdot k})$, for all $z_{i\cdot k}$ and $\tilde{z}_{i\cdot k}$, and consequently, as in the MAR mechanism, the intervals D_{ijk} are equal to $\mathbb{R}$.

4 Sparse estimation under multiple conditions and missingness

In this section, we develop sparse inference for the proposed model in the challenging setting where data are collected from different sub-populations, or generally experimental conditions, and where data are only partially observed either at the response or predictor levels. This is indeed the case of our applied setting, where there are three sub-populations of cells (Pre-MEP, E-MEP, and MK-MEP) and missingness both on X and on Y.

Maximum-likelihood estimators, defined as the maximizer the following average observed log-likelihood function

$$\bar{\ell}(\{\mathrm{Y}\}) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \log \int_{D_{i\cdot k}} \phi(z^{o}_{i\cdot k}, z^{no}_{i\cdot k}; \mathrm{Y}_k) \mathrm{d}z^{no}_{i\cdot k}, \tag{4.8}$$

are known to have a high variance when the sample sizes are not large enough. In addition, they do not allow to highlight similarities among the estimated networks. To overcome the inferential problems just described and to gain a more informative description of the data-generating process, in this paper we propose the following penalized estimator:

$$\{\widehat{\mathrm{Y}}\} = \arg\max_{\{\mathrm{Y}\}} \bar{\ell}(\{\mathrm{Y}\}) - \lambda P_{\alpha_1}(\{B\}) - \rho \widetilde{P}_{\alpha_2}(\{\Theta\}) - \nu \widetilde{P}_{\alpha_3}(\{\Omega\}), \tag{4.9}$$

where $\{B\} = \{B_1, \ldots, B_K\}$, $\{\Theta\} = \{\Theta_1 \ldots, \Theta_K\}$ and $\{\Omega\} = \{\Omega_1, \ldots, \Omega_K\}$ are the three dependence structures of interest, respectively, across the K sub-populations. We call this estimator the joint conditional glasso estimator (in short `jcglasso`), to distinguish it from the one of Danaher et al. (2014) which is typically referred to as joint glasso (in short `jglasso`). Aside from allowing the presence of missing data in the inferential procedure, the proposed estimator includes also the influence of covariates. This creates an additional layer of complexity, as both B_k and Ω_k vary across sub-populations, as with the precision matrix Θ_k.

The rationale for choosing the penalty functions in (4.9) is to select convex functions that encourage sparsity in each matrix as well as specific forms of similarity across the regression

coefficient matrices and the precision matrices. In particular, for the regression coefficient matrices, we propose the following sparse group lasso penalty function:

$$P_{\alpha_1}(\{B\}) = \alpha_1 \sum_{k=1}^{K}\sum_{h=1}^{p} \|\beta_{h\cdot k}\|_1 + (1-\alpha_1)\sum_{h=1}^{p}\left(\sum_{k=1}^{K}\|\beta_{h\cdot k}\|_2^2\right)^{1/2},$$

where $\alpha_1 \in [0, 1]$ is a parameter controlling the trade-off between a weighted lasso and a weighted group lasso penalty function. Although the previous penalty function is computational appealing, as we will discuss in Section 5.2, in principle other possible convex penalty functions could be used, such as the fused lasso-type penalty. The first penalty function is aimed at identifying zero regression coefficients, i.e. selecting the important associations between $X_{\cdot k}$ and $Y_{\cdot k}$, whereas the second penalty function encourages a similar pattern of sparsity of the regression coefficients across the different sub-populations. As for the precision matrices, both those corresponding to the dependence structure of $X_{\cdot k}$ and $Y_{\cdot k}$, we follow the proposal of Danaher et al. (2014). In particular, $\widetilde{P}_{\alpha_2}(\{\Theta\})$, and similarly $\widetilde{P}_{\alpha_3}(\{\Omega\})$, can take one of the following two forms:

$$\widetilde{P}_{\alpha_2}(\{\Theta\}) = \alpha_2 \sum_{k=1}^{K}\sum_{h\neq m}|\theta_{hmk}| + (1-\alpha_2)\sum_{k<k'}\sum_{h,m}|\theta_{hmk} - \theta_{hmk'}|, \tag{4.10}$$

or

$$\widetilde{P}_{\alpha_2}(\{\Theta\}) = \alpha_2 \sum_{k=1}^{K}\sum_{h\neq m}|\theta_{hmk}| + (1-\alpha_2)\sum_{h\neq m}\left(\sum_{k=1}^{K}\theta_{hmk}^2\right)^{1/2}, \tag{4.11}$$

where $\alpha_2 \in [0, 1]$. The fused lasso penalty (4.10) encourages a stronger form of similarity between the precision matrices, encouraging some entries of $\widehat{\Theta}_1, \ldots, \widehat{\Theta}_K$ to be identical across the K sub-populations. On the contrary, the group lasso penalty (4.11) encourages only a shared pattern of sparsity.

Although the computation of the proposed estimator may look infeasible due to the large number of penalty functions, in the next section we will show how the parametrization (3.4) and the factorization $\phi_z(z; \mathrm{Y}_k) = \phi_{y|x}(y \mid x; \beta_{0\cdot k}, B_k, \Theta_k)\phi_x(x; \mu_k, \Omega_k)$ lead to an extremely flexible and relatively easy to implement EM algorithm, where specific ADMM sub-routines can be run in parallel, enabling the study of high-dimensional datasets.

5 Computational aspects

The EM algorithm is grounded on the idea of repeating the expectation and maximization steps until a convergence criterion is met. We discuss in details these two steps.

5.1 The expectation step

Let $\{\widehat{\mathrm{Y}}\}$ the current estimate of the parameters of interest and use (3.4) to compute $\{\widehat{\Xi}\}$. The first step, also called E-Step, requires the calculation of the conditional expected value of the complete log-likelihood function. As our model assumes that, for each $k = 1, \ldots, K$, the random vector $Z_{\cdot k}$ follows a multivariate Gaussian distribution, which is a member of the exponential family, the E-Step reduces to the calculation of conditional expected values of the sufficient statistics (McLachlan & Krishnan, 2008). In our case, this results in imputed missing data (from the first moments) and imputed covariance matrices (from the second moments), which are then fed to the maximization step.

In particular, the missing values in the kth dataset are replaced with the following conditional expected values:

$$\hat{z}_{ijk} = E_{\hat{\psi}_k;\widehat{\Xi}_k}(Z_{ijk} \mid z_{i\cdot k} \in D_{i\cdot k}), \tag{5.12}$$

where $E_{\hat{\psi}_k;\widehat{\Xi}_k}(\cdot \mid z_{i\cdot k} \in D_{i\cdot k})$ is the expected value operator computed with respect to the multivariate Gaussian distribution truncated over the region $D_{i\cdot k}$. The resulting imputed dataset is denoted by $\widehat{Z}_k = (\widehat{X}_k; \widehat{Y}_k)$. Similarly, the matrix of second moments is computed as follows:

$$\widehat{C}_k = n_k^{-1} \sum_{i=1}^{n_k} E_{\hat{\psi}_k, \widehat{\Xi}_k}(Z_{i\cdot k} Z_{i\cdot k}^\top \mid z_{i\cdot k} \in D_{i\cdot k}) = \begin{pmatrix} \widehat{C}_{xx,k} & \widehat{C}_{xy,k} \\ \widehat{C}_{yx,k} & \widehat{C}_{yy,k} \end{pmatrix}.$$

This requires the computation of the moments of a multivariate truncated Gaussian distribution, with complex numerical algorithms for the calculation of the integral of the multivariate normal density function (see Genz & Bretz, 2002 for a review) that soon become computationally infeasible for the problem considered. Alternative efficient solutions are to either use a Monte Carlo method or to replace the mixed moments by the products of the conditional expectations, essentially calculating only conditional means and conditional variances. This second approach, proposed by Guo et al. (2015), was used successfully in a number of papers, e.g. (Augugliaro, Abbruzzo, et al., 2020; Behrouzi & Wit, 2019) among others, and will be considered also for the present paper.

5.2 The maximization step

The complete log-likelihood function resulting from the E-Step, called Q-function, represents the key element of the M-Step and it will be denoted by $Q(\{\psi\}, \{\Omega\}, \{B\}, \{\Theta\})$ to emphasize its dependence on the different sets of parameters.

Formally, the M-Step solves the following new maximization problem, obtained by replacing the marginal log-likelihood function in (4.9) with the Q-function:

$$\max_{\{Y\}} Q(\{\psi\}, \{\Omega\}, \{B\}, \{\Theta\}) - \lambda P_{\alpha_1}(\{B\}) - \rho \widetilde{P}_{\alpha_2}(\{\Theta\}) - \nu \widetilde{P}_{\alpha_3}(\{\Omega\}). \tag{5.13}$$

The particular structure of the Q-function for our problem allows to solve the maximization problem (5.13) through the solution of the following sequence of three sub-maximization problems:

$$\max_{\{\psi\}} Q(\{\psi\}, \{\widehat{\Omega}\}, \{\widehat{B}\}, \{\widehat{\Theta}\}), \tag{5.14}$$

$$\max_{\{\Omega\}} Q(\{\hat{\psi}\}, \{\Omega\}, \{\widehat{B}\}, \{\widehat{\Theta}\}) - \nu \widetilde{P}_{\alpha_3}(\{\Omega\}), \tag{5.15}$$

$$\max_{\{B\},\{\Theta\}} Q(\{\hat{\psi}\}, \{\widehat{\Omega}\}, \{B\}, \{\Theta\}) - \lambda P_{\alpha_1}(\{B\}) - \rho \widetilde{P}_{\alpha_2}(\{\Theta\}). \tag{5.16}$$

Each of these sub-problems involves only a set of parameters at once, while other parameters are held to their current estimates. Of particular notice is the fact that the last two sub-problems are unrelated to each other, in the sense that they can be written as separate optimization sub-routines involving only one set of parameters and the current estimates. Thus, they can be solved in parallel, improving the overall efficiency of the proposed algorithm and making it applicable to large datasets.

We start our discussion by considering sub-problem (5.14). Based on the current estimates of the parameters, the Q-function can be rewritten as:

$$Q(\{\psi\}, \{\widehat{\Omega}\}, \{\widehat{B}\}, \{\widehat{\Theta}\}) = -\sum_{k=1}^{K} f_k (\bar{z}_k - \psi_k)^\top \widehat{\Xi}_k (\bar{z}_k - \psi_k) + C,$$

where $\bar{z}_k = n_k^{-1} \widehat{Z}_k^\top 1_{n_k}$, $f_k = n_k/(2n)$ and C is a constant term. This then leads to the well-known optimal solution of sub-problem (5.14), given by:

$$\hat{\psi}_{x,k} = n_k^{-1} \widehat{X}_k^\top 1_{n_k} \quad \text{and} \quad \hat{\psi}_{y,k} = n_k^{-1} \widehat{Y}_k^\top 1_{n_k}, \quad k = 1, \ldots, K.$$

Given the new estimates $\{\hat{\psi}\}$, sub-problems (5.15) and (5.16) can be efficiently solved by factorizing the joint density function of $Z_{\cdot k}$ as a product of the conditional distribution of $Y_{\cdot k}$ and the marginal distribution of $X_{\cdot k}$. This step removes the problem of identifiability of the density of $Z_{\cdot k}$ due to the reparametrization (3.4). Indeed, thanks to this factorization, the Q-function admits the following additive structure:

$$\begin{aligned} Q(\{\hat{\psi}\}, \{\Omega\}, \{B\}, \{\Theta\}) &= Q_x(\{\hat{\psi}_x\}, \{\Omega\}) + Q_{y|x}(\{\hat{\psi}\}, \{B\}, \{\Theta\}) \\ &= \sum_{k=1}^{K} f_k \Big[\log\det\Omega_k - \mathrm{tr}\{\Omega_k \widehat{S}_{xx,k}\} \Big] + \sum_{k=1}^{K} f_k \Big[\log\det\Theta_k - \mathrm{tr}\{\Theta_k \widehat{S}_{y|x,k}(B_k)\} \Big], \end{aligned} \tag{5.17}$$

where $S_{xx,k} = \widehat{C}_{xx,k} - \hat{\psi}_{x,k}\hat{\psi}_{x,k}^{\top}$, $S_{yy,k} = \widehat{C}_{yy,k} - \hat{\psi}_{y,k}\hat{\psi}_{y,k}^{\top}$, $S_{yx,k} = \widehat{C}_{yx,k} - \hat{\psi}_{y,k}\hat{\psi}_{x,k}^{\top}$ and, finally, $\widehat{S}_{y|x,k}(B_k) = \widehat{S}_{yy,k} - \widehat{S}_{yx,k}B_k - B_k^{\top}\widehat{S}_{xy,k} + B_k^{\top}\widehat{S}_{xx,k}B_k$. The additivity of expression (5.17) justifies why problems (5.15) and (5.16) can be solved in parallel. Indeed, re-writing (5.15) using (5.17) and only the terms depending on $\{\Omega\}$, we have:

$$\{\widehat{\Omega}\} = \arg\max_{\{\Omega\}} \sum_{k=1}^{K} f_k \Big[\log\det\Omega_k - \mathrm{tr}\{\Omega_k \widehat{S}_{xx,k}\} \Big] - \nu \widetilde{P}_{\alpha_3}(\{\Omega\}), \tag{5.18}$$

which is equivalent to the definition of the `jglasso` (Danaher et al., 2014). So we solve this optimization using the efficient ADMM algorithm proposed in that paper.

Finally, considering the sub-problem (5.16) and using again (5.17), we have:

$$\max_{\{B\},\{\Theta\}} \sum_{k=1}^{K} f_k \Big[\log\det\Theta_k - \mathrm{tr}\{\Theta_k \widehat{S}_{y|x,k}(B_k)\} \Big] - \lambda P_{\alpha_1}(\{B\}) - \rho \widetilde{P}_{\alpha_2}(\{\Theta\}), \tag{5.19}$$

which represents an extension of the conditional glasso problem studied by various authors (Augugliaro, Sottile, et al., 2020; Augugliaro et al., 2023; Rothman et al., 2010; Yin & Li, 2011) to the case of multiple conditions. Since for any fixed $\{\hat{\psi}\}$, the function $Q_{y|x}(\{\hat{\psi}\}, \{B\}, \{\Theta\})$ is a bi-convex function of $\{B\}$ and $\{\Theta\}$, the maximization problem (5.19) can be carried out by repeating the two sub-steps of estimation of $\{B\}$ and estimation of $\{\Theta\}$, respectively, until a convergence criterion is met. In particular, given the current estimate of $\{\widehat{\Theta}\}$ the set $\{B\}$ can be estimated by solving the sub-problem:

$$\min_{\{B\}} \sum_{k=1}^{K} f_k \mathrm{tr}\{\widehat{\Theta}_k \widehat{S}_{y|x,k}(B_k)\} + \lambda P_{\alpha_1}(\{B\}), \tag{5.20}$$

whereas, given $\{\widehat{B}\}$, the set $\{\Theta\}$ is estimated by solving the sub-problem:

$$\max_{\{\Theta\}} \sum_{k=1}^{K} f_k \Big[\log\det\Theta_k - \mathrm{tr}\{\Theta_k \widehat{S}_{y|x,k}(\widehat{B}_k)\} \Big] - \rho \widetilde{P}_{\alpha_2}(\{\Theta\}). \tag{5.21}$$

Similar to problem (5.18), problems (5.20) and (5.21) can be solved via ADMM algorithms. Indeed, problem (5.21) corresponds to a `jglasso` where the standard empirical covariance for each condition is replaced by the conditional imputed covariance. On the other hand, problem (5.20) requires further development and is discussed separately.

Going then into the detail of problem (5.20), this could be solved, in principle, by the multi-lasso algorithm of Augugliaro, Sottile, et al. (2020). However, this algorithm tends to be unstable when the response vectors are highly correlated, as it is based on the idea of optimizing one column of each regression coefficient matrix at a time until a convergence criterion is met. This shortcoming

motivated us to develop a dedicated ADMM. In particular, the scaled augmented Lagrangian for sub-problem (5.20) is given by

$$L_\tau(\{B\}, \{\Gamma\}, \{U\}) = \sum_{k=1}^{K} f_k \mathrm{tr}\{\widehat{\Theta}_k \widehat{S}_{y|x,k}(B_k)\} + \lambda P_{\alpha_1}(\{\Gamma\}) + \frac{\tau}{2}\sum_{k=1}^{K} \|B_k - \Gamma_k + U_k\|_F^2 - \frac{\tau}{2}\sum_{k=1}^{K} \|U_k\|_F^2,$$

where U_k are dual matrices, $\tau > 0$ is a penalty parameter controlling the step size and $\|\cdot\|_F$ denotes the Frobenius norm. As described in Boyd et al. (2011), the ADMM algorithm is based on the idea of repeating the minimization of $L_\tau(\{B\}, \{\Gamma\}, \{U\})$ with respect to a set of parameters, while the remaining parameters are kept fixed to the previous values. Formally, the algorithm consists in repeating the following three steps:

$$\{B^{(i)}\} = \arg\min_{\{B\}} \textstyle\sum_{k=1}^{K} [f_k \mathrm{tr}\{\widehat{\Theta}_k \widehat{S}_{y|x,k}(B_k)\} + \frac{\tau}{2}\|B_k - \Gamma_k^{(i-1)} + U_k^{(i-1)}\|_F^2],$$
$$\{\Gamma^{(i)}\} = \arg\min_{\{\Gamma\}} \tfrac{\tau}{2}\textstyle\sum_{k=1}^{K} \|A_k - \Gamma_k\|_F^2 + \lambda P_{\alpha_1}(\{\Gamma\}),$$
$$\{U^{(i)}\} = \{U^{(i-1)}\} + \{B^{(i)}\} - \{\Gamma^{(i)}\},$$

until a convergence criterion is met. In the examples throughout the paper, we used $\tau = 2$ and declared convergence when $\sum_{k=1}^{K} \|B_k^{(i)} - B_k^{(i-1)}\|_1 / \sum_{k=1}^{K} \|B_k^{(i-1)}\|_1 < 10^{-5}$. Detailed derivations of the algorithm and pseudo-codes are reported in the online supplementary material.

Before closing this section, we discuss the convergence of the proposed algorithm. First, we recall that standard theory on the EM algorithm tells us that the objective function increases at each iteration of the EM algorithm. Although this does not imply that the solution of the EM algorithm is the global optimal one, this is true under fairly general conditions (see McLachlan & Krishnan, 2008 for more details). Another element that could influence the convergence of the proposed algorithm is the bi-convexity of the function $Q_{y|x}(\{\hat{\psi}\}, \{B\}, \{\Theta\})$. This aspect was studied by Yin and Li (2011) in the context of conditional glasso estimators. The authors prove that the two-steps algorithm discussed above always converges to a stationary point. In the classical setting, where the sample size is large enough, the global optimality is always assured, while, in a high-dimensional setting, the convergence properties closely depend on the ratio between sample size and the number of nonzero estimates or, in other terms, on the values of the tuning parameters used. When λ and ρ are large enough with respect to the sample size, the algorithm converges to the unique stationary point. In contrast, when the two tuning parameters are too close to zero, there are potentially many stationary points due to the high-dimensional nature of the parameter space. This behaviour is not surprising and is shared by many algorithms developed for inference in the case of high dimensionality. In that regard, a key role will be played by the method used to select the tuning parameters, which will be discussed in the next section.

6 Tuning parameters selection

The full inference is conducted for a selection of tuning parameters, namely α_1, α_2, α_3, that allow to balance the two convex penalties, and λ, ρ, ν, that control the degree of sparsity of the solution. For each combination of the six tuning parameters, the optimal model can be selected using the Bayesian Information Criterion (BIC), as suggested by many authors (Li et al., 2012; Städler & Bühlmann, 2012; Yin & Li, 2011). For the proposed model, this is defined by

$$\mathrm{BIC} = -2n\bar{\ell}(\{\widehat{\Upsilon}\}) + \mathrm{df}\sum_{k=1}^{K} \log n_k,$$

where df denotes the number of distinct nonzero estimated parameters in the regression matrices and precision matrices, and $\{\widehat{\Upsilon}\}$ is the MLE of the parameters constrained to the sparsity pattern of the solution at the given tuning parameters. Two problems appear: the first one is that the average

log-likelihood in the BIC definition, defined in (4.8), is not a direct output of the EM algorithm and its calculation adds a significant computational burden to the procedure; the second one is the large number of tuning parameters.

With regards to the first problem, following Ibrahim et al. (2008) and Augugliaro, Sottile, et al. (2020); Augugliaro et al. (2023), we replace the exact log-likelihood function with the Q-function used in the M-Step of the main algorithm, i.e. we use the following approximation:

$$\begin{aligned}\overline{\mathrm{BIC}}_z &= -2nQ(\{\widehat{\Upsilon}\}) + \mathrm{df}\sum_{k=1}^{K}\log n_k \\ &= -2nQ_x(\{\hat{\psi}_x\}, \{\widehat{\Omega}\}) + \mathrm{df}_x\sum_{k=1}^{K}\log n_k - 2nQ_{y|x}(\{\hat{\psi}_y\}, \{\widehat{B}\}, \{\widehat{\Theta}\}) + \mathrm{df}_{y|x}\sum_{k=1}^{K}\log n_k \\ &= \overline{\mathrm{BIC}}_x + \overline{\mathrm{BIC}}_{y|x}\end{aligned} \tag{6.22}$$

where df_x and $\mathrm{df}_{y|x}$ correspond to the number of distinct nonzero parameters in $\{\widehat{\Omega}\}$ and in $(\{\widehat{B}\}, \{\widehat{\Theta}\})$, respectively. Expression (6.22) is computationally efficient as the Q-function is easily obtained as a byproduct of the EM algorithm. Moreover, it is worth noting that $\overline{\mathrm{BIC}}_z$ is equal to the sum of two specific measure of goodness-of-fit, denoted by $\overline{\mathrm{BIC}}_x$ and $\overline{\mathrm{BIC}}_{y|x}$, respectively, that will play a central role in the sub-selection procedures discussed below.

With regards to the second problem, we have devised an efficient computational strategy on how to streamline the procedure. This is mainly based on two observations, which are supported by the results of the simulation study that will be discussed in Section 7.2.

The first observation is that the α parameters have little effect on the selection procedure of the optimal values of the remaining shrinkage tuning parameters. This implies that we can split the entire selection procedure into a sequence of steps. First, given suitable α starting values, we select the optimal values of the remaining shrinkage parameters. Then, given these, we select the optimal values of the mixing parameters α. It is worth noting that such a procedure can be further simplified if, as suggested by Danaher et al. (2014), one is able to fix some a priori known α-values, since in that case one has to select only the shrinkage parameters.

The second observation is that the behavior of the estimates of $\{\widehat{B}\}$ and $\{\widehat{\Theta}\}$ is largely unaffected by the estimates of the parameters of the marginal distribution of X. Consequently, given the values of the mixing parameters α, the selection procedure of the shrinkage parameters ν, λ, and ρ can be split into a sub-selection procedure aimed at selecting ν and a sub-selection procedure for selecting λ and ρ. While the first one is a standard problem in the glasso literature and can be solved by evaluating $\overline{\mathrm{BIC}}_x$ over a sequence of ν-values, the second one is more challenging as it requires the evaluation of the $\overline{\mathrm{BIC}}_{y|x}$ measure over a two-dimensional grid of the tuning parameters, say $\Lambda = \{\lambda_{\min}, \ldots, \lambda_{\max}\} \times P = \{\rho_{\min}, \ldots, \rho_{\max}\}$. When the sample size is large enough, one can let $\lambda_{\min}$ and $\rho_{\min}$ equal to zero, so that the maximum-likelihood estimate is one of the points belonging to the coefficient path. On the other hand, in a high-dimensional setting, one needs to use two values small enough to avoid the instability of the model. With respect to the largest values of the two tuning parameters, combining the results of Theorems 1 and 2 of Danaher et al. (2014) and of Theorem 3 of Augugliaro, Sottile, et al. (2020) the formulae for computing $\lambda_{\max}$ and $\rho_{\max}$ are given, respectively, by

$$\lambda_{\max} = \left\| \left\{ \sum_{k=1}^{K} \left(f_k \widehat{S}_{xy,k} \widehat{\Theta}_k \right)_+^2 \right\}^{1/2} \right\|_\infty,$$

$$\rho_{\max} = \begin{cases} \max\left\{ \left\| f_1 \widehat{S}_{yy,1} \right\|_\infty^-, \ldots, \left\| f_K \widehat{S}_{yy,K} \right\|_\infty^- \right\} & \text{for the fused lasso penalty 4.10} \\ \left\| \left\{ \sum_{k=1}^{K} \left(f_k \widehat{S}_{yy,k} \right)_+^2 \right\}^{1/2} \right\|_\infty^- & \text{for the group lasso penalty 4.11} \end{cases}$$

where $(\cdot)_+^2$ indicates the square of the maximum between zero and the argument and $\|\cdot\|_\infty^-$ indicates the maximum element in absolute value of the upper triangular part of its argument. Once $\lambda_{\max}$ and

$\rho_{\max}$ are calculated, the entire coefficient path can be computed over a two-dimensional grid $\Lambda \times P$ using the estimate obtained for a given pair (λ, ρ) as warm starts for fitting the next model.

7 Simulation study

In this section, we evaluate the proposed `jcglasso` method. Firstly, we study its performance under different settings, aimed in particular at evaluating the tuning parameter selection method proposed in Section 6. Secondly, we compare our approach to two existing methods.

7.1 Data-generating process

To simulate data under different conditions, the following setting is used. We set the number of groups to $K = 2$, the sample size per group to $n_k = 100$, the number of response variables $p \in \{50, 200\}$ and the number of covariates $q = 50$. Responses and covariates are simulated using a multivariate Gaussian distribution, whose precision matrices $\{\Theta\}$ and $\{\Omega\}$, respectively, are generated from a sparse random graph structure and with a high sharing between the two sub-populations. In particular, using the generator function implemented in the `BDgraph` package (Mohammadi & Wit, 2019), we simulate Θ_1 and Ω_1 in the first sub-population in such a way that the probability of a nonzero entry is equal to 0.01 and 0.05, respectively. The matrices Θ_2 and Ω_2 from the second sub-population are constructed by randomly drawing 50% of the nonzero entries in Θ_1 and Ω_1, respectively. A similar strategy is used to simulate sparse regression coefficient matrices with a 50% sharing between the two sub-populations. First, the matrix B_1 in sub-population 1 is generated so that for each response variable only two regression coefficients are nonzero, with their value sampled from a uniform distribution on the interval [0.50, 0.70]. Then B_2 is constructed to share 50% of the nonzero regression coefficients in B_1. Figure 2 shows an image of the matrices in a simple setting ($q = 25$), revealing their high sparsity and sharing between sub-populations. Finally, responses and covariates are subject to 25% missingness at random, and for each of them the probability of missing is set to 0.25.

The estimators are evaluated in terms of their ability to recover the true sparse structures and their accuracy in estimating the true precision and regression coefficient values. In particular, we evaluate the ability of the method in identifying the nonzero entries of the matrices $\{\widehat{\Theta}\}$, and similarly $\{\widehat{\Omega}\}$ and $\{\widehat{B}\}$, with the area under the precision-recall curve (AUC-PR). One point of the curve corresponds to a given choice of the tuning parameters and, for the corresponding $\{\widehat{\Theta}\}$, it is given by

$$\text{Precision} = \frac{\text{number of } \hat{\theta}_{hm,k} \neq 0 \text{ and } \theta_{hm,k} \neq 0}{\text{number of } \hat{\theta}_{hm,k} \neq 0},$$

$$\text{Recall} = \frac{\text{number of } \hat{\theta}_{hm,k} \neq 0 \text{ and } \theta_{hm,k} \neq 0}{\text{number of } \theta_{hm,k} \neq 0},$$

where $k = 1, \ldots, K$. We evaluate the accuracy in the parameter estimates with the mean squared error (MSE) which is defined, for a given choice of tuning parameters, by

$$\text{MSE} = \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \| \widehat{\Theta}_k - \Theta_k \|_F^2.$$

Given the large number of tuning parameters, the general strategy will be to evaluate the measures on $\{\widehat{B}\}$ by first obtaining $\{\widehat{\Theta}\}$ using a decreasing sequence of ρ-values, namely $\rho/\rho_{\max} \in \{1.00, 0.75, 0.50, 0.25\}$ and then exploring a grid of values of the tuning parameter λ by fixing the ratio $\lambda/\lambda_{\max} \in \{1.00, 0.90, \ldots, 0.20, 0.10\}$. Conversely, to obtain results on $\{\widehat{\Theta}\}$, the role between the tuning parameters λ and ρ is reversed, while keeping the other considerations unchanged. In all simulations, we use 50 simulated samples. Details specific to each simulation are given in the following subsections.

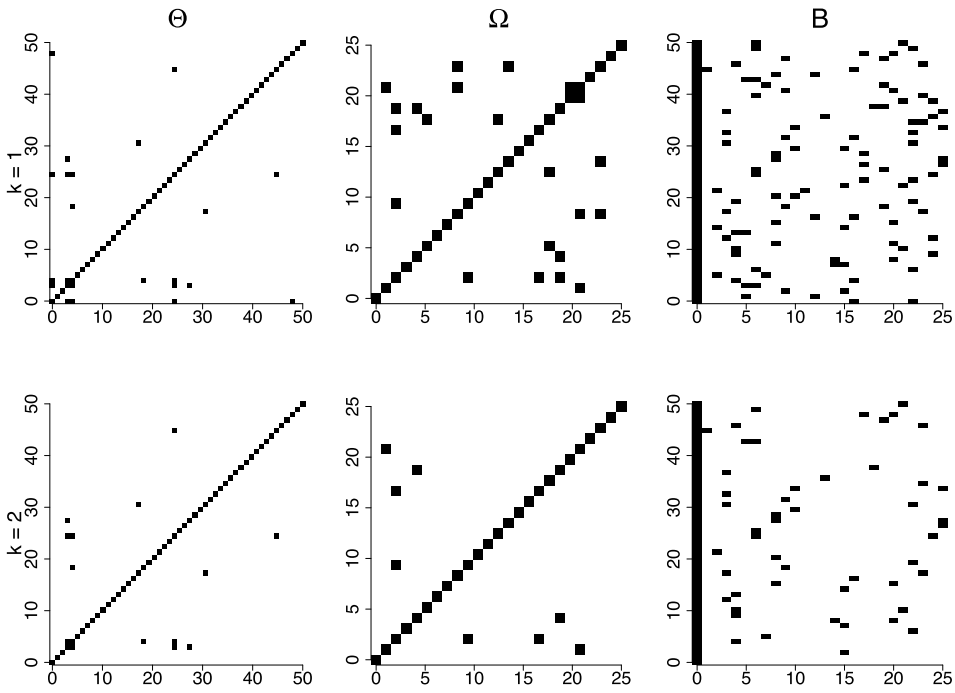

Figure 2. Random sparse structures of Θ_k, Ω_k, and B_k for $p = 50$, $q = 25$, and $k = 1, 2$. The probability of a nonzero entry is set to 0.01 in Θ_1 and 0.05 in Ω_1, while B_1 is generated as described in the manuscript. For the second condition, we randomly draw 50% of nonzero entries from the matrices B_1, Θ_1, and Ω_1, respectively.

7.2 Tuning parameter selection

In this subsection, we measure the performance of the proposed method under several settings chosen to evaluate the procedure we have devised in Section 6. We investigate how robust the selection procedure is for λ and ρ, given different choices of the tuning parameters v and α. In particular, we study how the MSE curves, taken as a proxy for guiding the choice of λ and ρ, change when varying the α values according to $\alpha_1 = \alpha_2 = \alpha_3 \in \{0.25, 0.50, 0.75\}$ and the tuning parameter $v \in \{0, v_{\text{BIC}}, v_{\max}\}$, where $v_{\max}$ is the maximum value such that $\{\widehat{\Omega}\}$ is diagonal and v_{BIC} is the optimal value obtained by minimizing the $\overline{\text{BIC}}_x$ criterion, most likely between zero and the maximum.

Looking at the MSE curves reported in Figure 3, the shapes are almost the same across the various settings, with a similar distribution of the minimum over each tuning parameter value sequence. The same behaviour is shown by the AUC-PR curves reported in Figure 4. Similar results have been found using denser random graph structures or block diagonal structures for $\{\Omega\}$ and $\{\Theta\}$, and are reported in the online supplementary material. Overall, these results indicate that the choice of the optimal tuning parameters λ and ρ is largely unaffected by the choice of the remaining tuning parameters. This supports the idea of dividing the selection procedure into three steps. The first step involves fixing the α mixing parameters to some a priori values, such as 0.25 or 0.75, depending on whether more weight is to be given to the group lasso penalty or vice versa. If these values are unknown, starting values can be used, e.g. $\alpha_1 = \alpha_2 = \alpha_3 = 0.5$ to provide an equal weight between the penalties. Then, the second step involves optimizing the tuning parameter v related to sub-problem (5.15) and the parameters λ and ρ related to sub-problem (5.16). Finally, if α is not known a priori, one can search again for the α values that minimize the chosen model selection criterion. This procedure does not compromise the overall selection process,

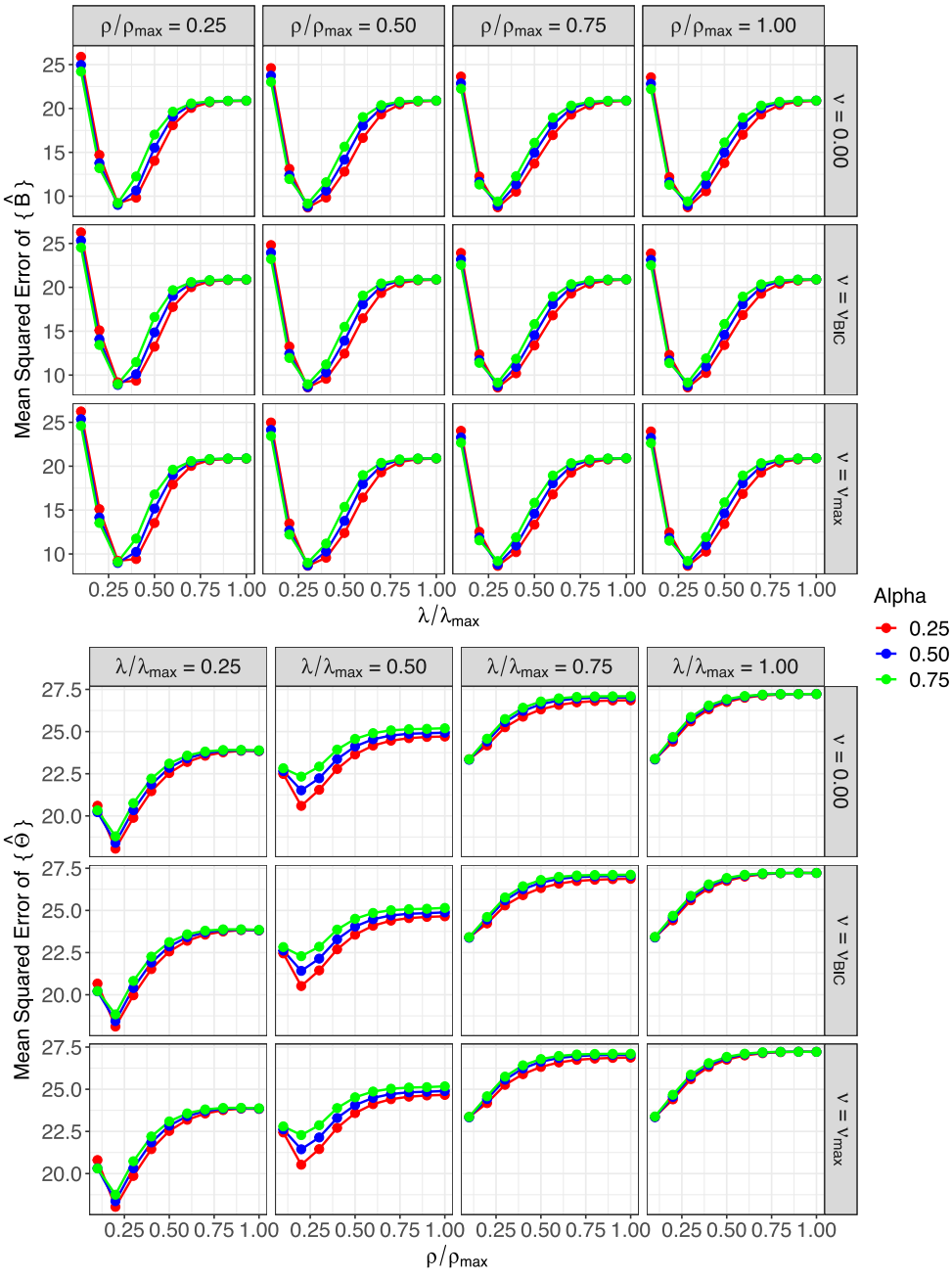

Figure 3. Simulation study with $p = 200$, $q = 50$, $n_k = 100$, 25% of response variables and covariates with a 0.25 probability of missing-at-random. The three lines in each plot correspond to `jcglasso` with $\alpha_1 = \alpha_2 = \alpha_3 \in \{0.25, 0.50, 0.75\}$. The two blocks of panels refer to the MSE curve for $\{\widehat{B}\}$ and $\{\widehat{\Theta}\}$. In each block, the rows show results for three different values of the parameter ν, namely $\nu = 0$, ν chosen by minimizing the Bayesian Information criterion and $\nu = \nu_{\max}$. The MSE curves of $\{\widehat{B}\}$ are computed over the ratio $\lambda/\lambda_{\max}$, where the structure of the precision matrix $\{\widehat{\Theta}\}$ is fixed by the ratio $\rho/\rho_{\max}$ (first block, by column). The MSE curves of $\{\widehat{\Theta}\}$ are evaluated over the ratio $\rho/\rho_{\max}$ fixing the structure of the regression coefficient matrix $\{\widehat{B}\}$ by the ratio $\lambda/\lambda_{\max}$ (second block, by column).

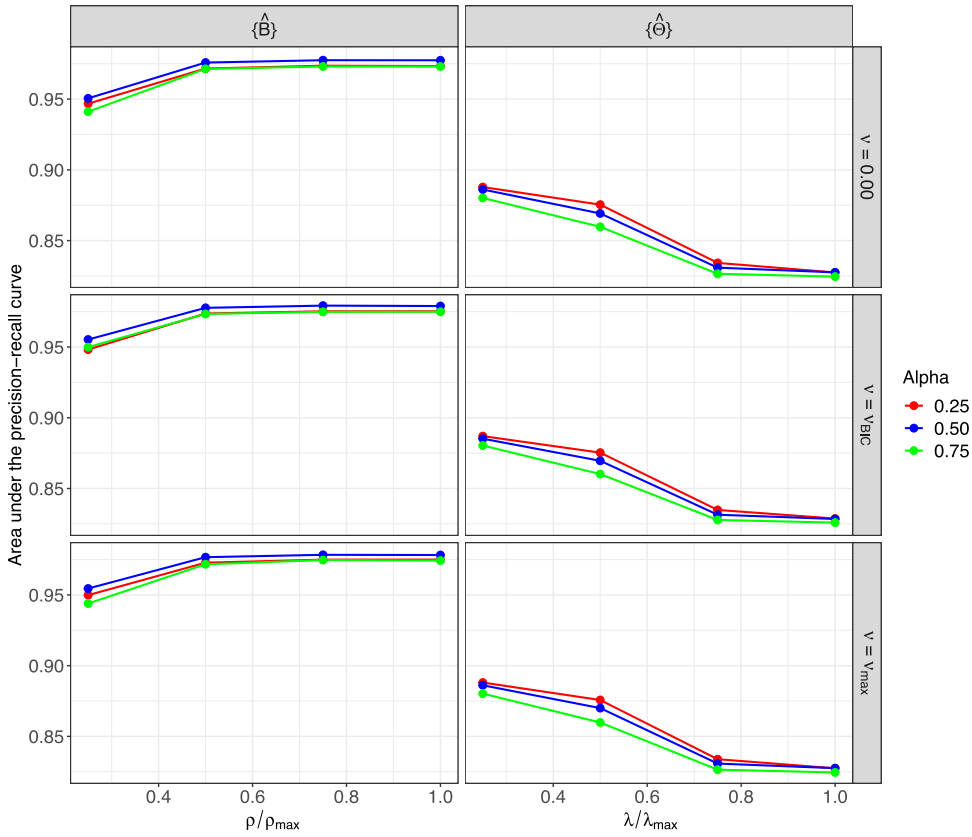

Figure 4. Simulation study with $p = 200$, $q = 50$, $n_k = 100$, 25% of response variables and covariates with a 0.25 probability of missing-at-random. The three lines in each plot correspond to `jcglasso` with $\alpha_1 = \alpha_2 = \alpha_3 \in \{0.25, 0.50, 0.75\}$. The panels refer to the area under the precision-recall curve (AUC-PR) curve for $\{\widehat{B}\}$ and $\{\widehat{\Theta}\}$. The rows show results for 3 different values of the parameter v, namely $v = 0$, v chosen by minimizing the Bayesian Information criterion and $v = v_{\max}$. The AUC-PR curves of $\{\widehat{B}\}$ (first column) are computed over the ratio $\lambda/\lambda_{\max}$, where in each panel each dot correspond to a different ratio $\rho/\rho_{\max}$. The AUC-PR curves of $\{\widehat{\Theta}\}$ (second column) are evaluated over the ratio $\rho/\rho_{\max}$, where in each panel each dot correspond to a different ratio $\lambda/\lambda_{\max}$.

confirming the considerations made in Section 6. Finally, we would like to note that, as also mentioned in Augugliaro, Sottile, et al. (2020); Augugliaro et al. (2023), the estimation of $\{\widehat{B}\}$ is not particularly affected by the sparse structure of $\{\widehat{\Theta}\}$, as shown in the upper panels of Figure 3. On the contrary, the results for $\{\widehat{\Theta}\}$ seem to depend strongly on the structure of $\{\widehat{B}\}$.

7.3 Evaluating the performance and comparison with existing methods

In this second simulation study, the behaviour of the proposed `jcglasso` is compared with the `cglasso` of Augugliaro et al. (2023) and the `jmmle` of Majumdar and Michailidis (2022). We underlie that the two competitors are proposed for a setting that is different from the one studied in our paper. In particular, the method of Augugliaro et al. (2023) allows data imputation under an MAR and censoring mechanism via an efficient EM algorithm, but does not take into account neither similarities between dependence structures across conditions, nor a sparse structure for the precision matrix of the predictors or an imputation mechanism for the covariates. On the other hand, the approach of Majumdar and Michailidis (2022) is proposed only in the case where the dataset is fully observed and it does not estimate the precision matrix Ω, but it allows to separate the estimation of $\{B\}$ and $\{\Theta\}$ and it imposes a shared structure over the K populations.

To be able to use the existing approaches, we proceed as follows:

- for `cglasso`, we estimate separate models for each sub-population. We consider the approach of Augugliaro, Abbruzzo, et al. (2020) to obtain the precision matrices $\{\Omega\}$, by selecting the minimum value of the MSE curve, and to impute the matrix of covariates X. The response Y_k, with potential missing entries, and the imputed matrix of covariates, are then used to estimate B_k and Θ_k using the model proposed in Augugliaro, Sottile, et al. (2020); Augugliaro et al. (2023);
- for `jmmle`, we first impute both responses and covariates according to a random forest, by using the `missForest` R package (Stekhoven & Buehlmann, 2012). Then, we estimate the precision matrices $\{\Omega\}$ by the method proposed in Ma and Michailidis (2016). Finally, we estimate $\{B\}$ and $\{\Theta\}$ accordingly to the model proposed in Majumdar and Michailidis (2022).

Table 1 presents the AUC and MSE values for our proposed approach and the two competitors across all conditions. For our methods, and similarly for the two competitors, we report the minimum value achieved over the sequence of tuning parameters ν, λ and ρ for different values of α. The online supplementary material include further simulation results using different graph structures and a comparison of the computational time between the three methods. The results demonstrate similar performance among the methods in terms of graph recovery, but a better performance of the proposed approach in terms of parameter estimation. Moreover, the online supplementary material show how the proposed approach is associated to a much lower computational time. It is important to note that `jmmle` and `cglasso` do not solve the entire problem, but require custom procedures, as described above. In contrast, the proposed approach provides a unified framework capable of capturing various sparse structures among sub-populations, while handling partially observed data.

8 Inferring the human haematopoietic system

The proposed `jcglasso` model is now used to study the data generated by Psaila et al. (2016), in order to gain insight into the process of blood cell formation described in the Introduction. Furthermore, common pathways between the three sub-populations of cells are expected, and the main interest is to identify their key differences. We used the strategy identified and evaluated in the previous sections to select the tuning parameters. First, we set the mixing parameters α to 0.5 and search for the best tuning parameter ν that infers the sparsity structure of the 40 receptors for each condition ($\{\widehat{\Omega}\}$), followed by the selection of the pair (λ, ρ) that infers the regulatory networks of the 34 nuclear activities ($\{\widehat{\Theta}\}$) and the associations between them and the membrane receptor activities ($\{\widehat{B}\}$), and finally we address the problem of selecting the tuning parameters α in the set $\{0.25, 0.50, 0.75\}$. We use $\overline{\text{BIC}}$ for all tuning parameters, setting an equally spaced sequence of 50 values for the selection of ν and a 10×10 grid of values for the selection of λ and ρ, given the previously identified optimal ($\{\widehat{\Omega}\}$). Figure 5 shows the $\overline{\text{BIC}}$ values along the solution path and the optimal values that result from this analysis ($\lambda = 0.074$, $\rho = 0.060$, and $\nu = 0.065$). Finally, we set $\alpha_1 = \alpha_2 = \alpha_3 = 0.75$, which is the value that minimizes the chosen criterion.

Looking now at the optimal networks, the number of unique nonzeros on the matrices associated to each population ($\widehat{\Omega}$, $\widehat{\Theta}$, $\widehat{B}$) is $\text{df}_{\text{PRE}} = 209$ (7.74%), $\text{df}_{\text{E}} = 199$ (7.37%) and $\text{df}_{\text{MK}} = 222$ (8.22%), respectively (see Table 2 for more details on the number of links at each level). Figure 6 (left) reports the number of edges that are common between, or specific within, each population. The right figure focuses only on those edges that have been experimentally validated. These are found using the 'connect' tool of the Ingenuity Pathways Analysis software (Qiagen IPA, October 2021; Krämer et al., 2014). Figure 7 focuses on these selected links, and distinguishes the links according to whether they refer to associations between the two levels (red arrows, corresponding to nonzeros in B) or to regulatory networks within the membrane receptors or nuclear factors (undirected edges, corresponding to nonzeros in Ω or Θ). The latter are further distinguished between interactions that are common to at least two populations (in grey) versus those that are specific to one population (in green). A further characterization of an edge is in terms of a positive partial correlation (solid line) versus a negative partial correlation (dashed line).

From a biological point of view, the analysis revealed a shared network between all the differentiation states studied. This network consists of proteins that are essential for the maintenance of

Table 1. Simulation study with $p = 200$, $q = 50$, $n_k = 100$, 25% of covariates with a 0.25 probability of missing-at-random, according to the sparse structure reported in Figure 2

			jmmle	cglasso	jcglasso		
					$\alpha = 0.25$	$\alpha = 0.50$	$\alpha = 0.75$
$\widehat{\Omega}$	AUC	$k = 1$	0.901	0.891	0.899	0.898	0.898
		$k = 2$	0.947	0.935	0.943	0.942	0.939
	MSE	$k = 1$	6.778	4.117	4.226	4.246	4.261
		$k = 2$	2.877	2.486	2.445	2.451	2.456
$\widehat{B}$	AUC	$k = 1$	0.981	0.970	0.960	0.959	0.951
		$k = 2$	0.903	0.984	0.967	0.978	0.975
	MSE	$k = 1$	5.476	5.482	5.05	5.023	5.090
		$k = 2$	3.922	4.226	4.134	3.958	3.957
$\widehat{\Theta}$	AUC	$k = 1$	0.850	0.844	0.848	0.850	0.846
		$k = 2$	0.919	0.909	0.921	0.920	0.914
	MSE	$k = 1$	12.984	10.692	10.876	10.908	10.981
		$k = 2$	8.405	7.165	7.554	7.506	7.602

Note. The columns show the compared methods `jmmle`, `jcglasso` with $\alpha_1 = \alpha_2 = \alpha_3 \in \{0.25, 0.50, 0.75\}$ and `cglasso`, and the rows show the statistics used to evaluate the performance of $\{\widehat{\Omega}\}$, $\{\widehat{B}\}$, and $\{\widehat{\Theta}\}$ in each sub-population k. AUC = area under curve.

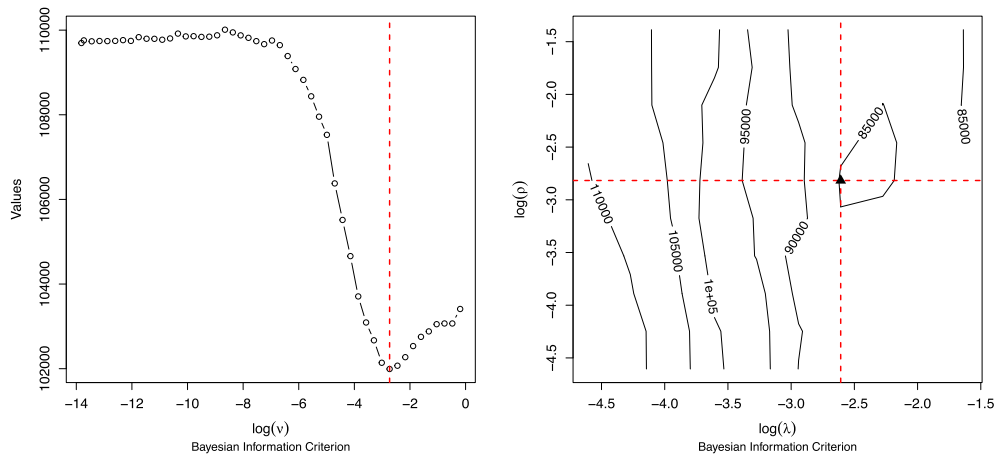

Figure 5. Selection of the (v, λ, ρ) tuning parameters on the real data analysis (after setting $\alpha_1 = \alpha_2 = \alpha_3 = 0.75$). Left: Bayesian Information $\overline{\text{BIC}}$ across a path of 50 v values; right: $\overline{\text{BIC}}$ contour lines on a two-dimensional grid of 100 (λ, ρ) values given the $\{\widehat{\Omega}\}$ estimated at the optimal v. BIC, Bayesian Information criterion.

haematopoietic stem cell properties. Table 3 shows the ranking of putative associations in terms of estimated partial correlations coefficients, across the three conditions. Specifically, *TAL1*, *RUNX1*, and *GATA2* act as master regulators of haematopoiesis (Vagapova et al., 2018) and are central to all three conditions studied, while *GATA1* appears to be recruited to the network only in the more differentiated forms. Cell cycle regulation in blood stem cells is highly dependent on the interaction of *MYB* with *GATA2* and thus with *CDK4*, *CDK6*, and *CDKN1B* (Matsumoto & Nakayama, 2013). It is noteworthy that while the *MYB-GATA2* network remains critical in pre-MEP and the most differentiated MK-MEP, it appears to be relatively less important in the E-MEP lineage. The role of *CDK4* and *CDK6* (see Table 3) is interesting, as both proteins function as cell cycle regulators and share many activities. However, they are differentially recruited in the

Table 2. Number of unique nonzeros (and density) of the optimal networks associated to each population ($\widehat{\Omega}$, $\widehat{\Theta}$, $\widehat{B}$)

		Pre-MEP		E-MEP		MK-MEP	
		#edges	Density (%)	#edges	Density (%)	#edges	Density (%)
$\widehat{\Omega}$	All	94	12.1	101	13.0	123	15.8
	Validated	3	0.4	2	0.3	8	1.0
$\widehat{B}$	All	46	3.4	55	4.0	37	2.7
	Validated	12	0.9	9	0.7	6	0.4
$\widehat{\Theta}$	All	69	12.3	43	7.7	62	11.1
	Validated	28	5.0	16	2.9	26	4.6
Overall	All	209	7.7	199	7.4	222	8.2
	Validated	43	1.6	27	1.0	40	1.5

Note. For each matrix, the first row refers to the overall number of links and the second focuses on those that have been experimentally validated. Pre-MEP = pre-megakaryocyte-erythroid progenitor (MEP); E-MEP = erythroid-MEP; MK-MEP = megakaryocyte-MEP.

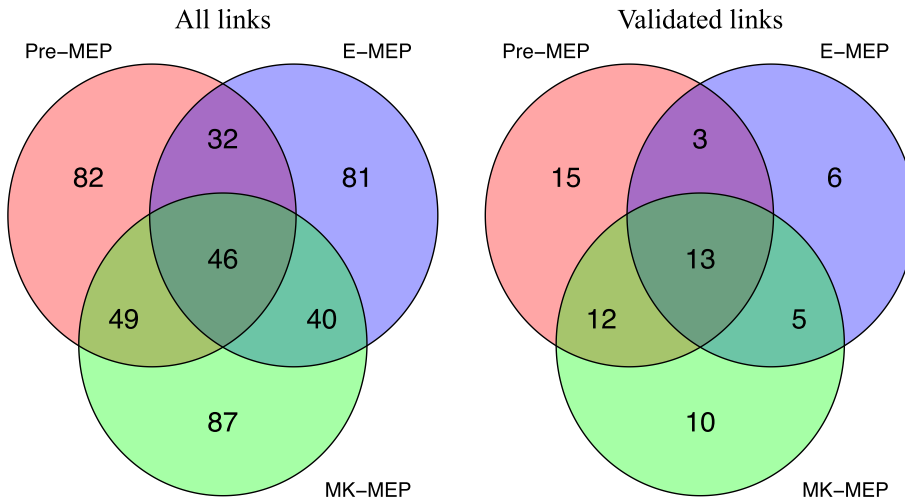

Figure 6. Venn diagrams reporting the number of edges common and specific to the three groups of megakaryocytes (left) and the number of those that have been experimentally validated (right). The latter is found using the IPA software.

three networks studied, suggesting that their activities are only partially congruent, as previously suggested in the literature (Scheicher et al., 2015).

Examining the specific characteristics of each sub-population, *CD44* is particularly prominent in the pre-MEP network and is associated with the most primitive state, as suggested in the literature (Shin et al., 2014). Furthermore, a more detailed analysis using the data presented in Table 3 shows that its positive partial correlation with *CD9* (0.482) and negative partial correlation with *CD71/TFRC* (−0.223), which are markers of more differentiated states, is consistent with the versatile role of pre-MEP cells. In turn, we can speculate that *CD71/TFRC* may regulate *RUNX1* (1.038). Overall, the central role of *CD44* allows us to present a scenario in which *CD44* is upregulated in pre-MEP and negatively correlated with *CD71/TFRC*, which is activated in the more differentiated form and promotes the activity of *RUNX1*, as in E-MEP (0.310).

The E-MEP shows a simplified network compared to the Pre-MEP. This is consistent with the model of transformation of the Pre-MEP into the E-MEP. In particular, some of the network nodes have different connections, as in the case of *CD117/KIT*. It is worth noting that this change in the

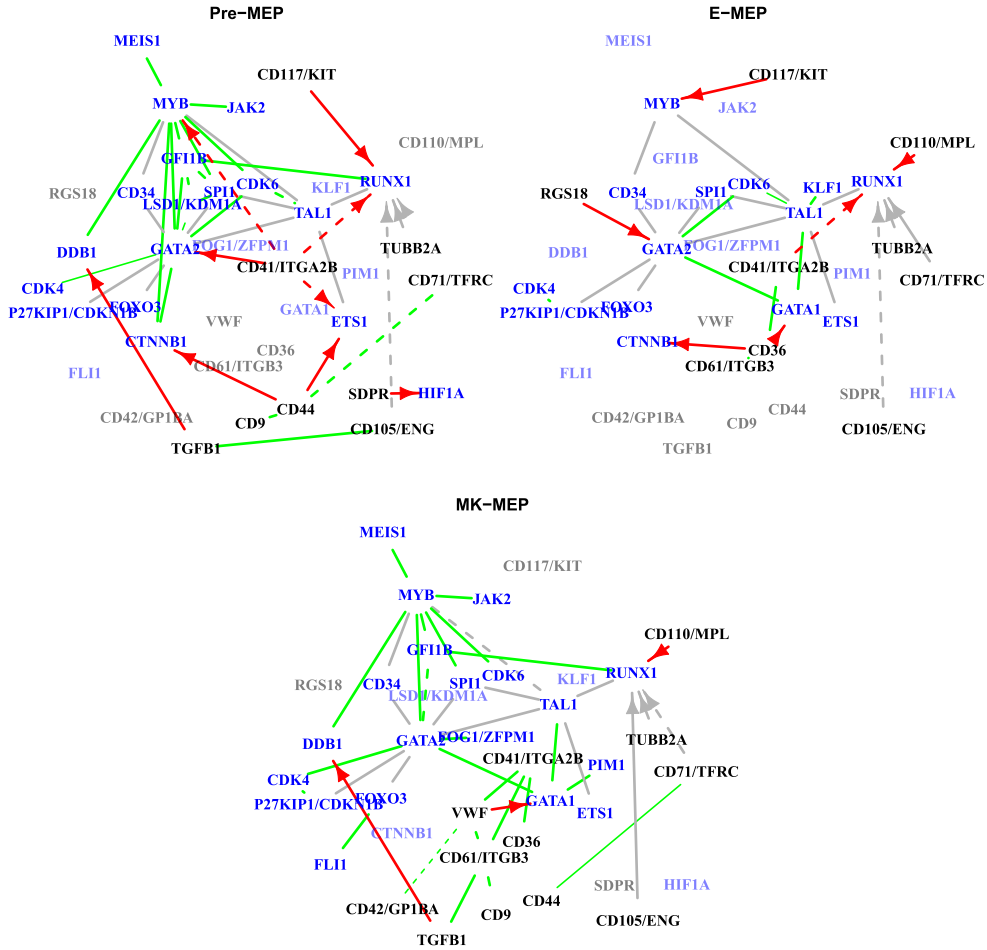

Figure 7. Regulatory networks of the nuclear and membrane factors across the three sub-populations of cells. In each graph, blue, black, and transparent nodes refer to nuclear factors, membrane receptors and isolated nodes, respectively. Red arrows indicate edges from membrane receptors to nuclear factors; grey links indicate edges shared among the three populations; green links refer to specialized ones. Solid and dashed lines indicate positive and negative partial correlations, respectively. In contrast, the thickness of the line refers to the partial correlation value standardized to the maximum value inside the associated network.

network could be related to the fact that low expression of *c-Kit* is associated with more immature forms (Shin et al., 2014). In addition, *CD36* has a unique role within E-MEP. Although *CD36* has traditionally been associated with the megakaryocytic lineage, recent studies suggest that it has a specific effect on erythropoiesis (Meng et al., 2023). Additionally, *CD36* appears to have associations within the membrane bound factors, as it specifically binds to *ITGA2B* (0.133) and to a lesser extent to *ITGB2* (0.024). This example illustrates how the approach outlined in this paper can reveal novel features and prove beneficial in biomedical research. This is especially true when analysing very complex scenarios, as this analysis allows focusing on the strongest links, which are likely to be the most relevant in the network, and progressively analysing the weaker ones as the analysis requires. Finally, *CD110/MPL* and *CD71/TFRC* are markers of partial binding, but surprisingly remain present in both E-MEP and MK-MEP lines, indicating their immature nature. However, the highly specific megakaryocytic marker *VWF* is exclusively present in the MK-MEP network, indicating its central role in MK-MEP-specific commitment. *TGFB1* is not present in the E-MEP network, and is known to prevent differentiation towards the erythroid lineage in favour of the megakaryocytic lineage (Blank & Karlsson, 2015). Overall, the results

Table 3. Regulatory networks of the nuclear and membrane factors in the three sub-populations of cells

	Pre-MEP			E-MEP			MK-MEP		
	from	to	PC	from	to	PC	from	to	PC
$\widehat{\Omega}$	CD44	CD9	0.482†	CD36	CD41#	0.133†	CD41#	CD61#	0.243†
	CD44	CD71#	−0.223†	CD36	CD61#	0.024†	CD61#	VWF	0.237†
	CD105#	TGFB1	0.087†				CD36	CD41#	0.198†
							CD61#	TGFB1	0.189†
							CD41#	VWF	0.134†
							CD61#	CD9	−0.028†
							CD44	CD71#	0.008†
							CD42#	VWF	−0.005†
$\widehat{B}$	CD71#	RUNX1	1.038	CD110#	RUNX1	0.619†	CD110#	RUNX1	0.300†
	CD117#	RUNX1	0.586†	CD36	CTNNB1	0.590†	CD105#	RUNX1	0.224
	SDPR	HIF1A	0.535†	CD117#	MYB	0.531†	VWF	GATA1	0.181†
	CD105#	RUNX1	−0.503	RGS18	GATA2	0.452†	TGFB1	DDB1	0.138†
	CD44	CTNNB1	0.333†	CD41#	RUNX1	−0.362†	TUBB2A	RUNX1	0.054
	CD41#	RUNX1	−0.312†	CD71#	RUNX1	0.310	CD71#	RUNX1	−0.040
	CD41#	ETS1	−0.217†	CD105#	RUNX1	−0.206			
	CD41#	GATA2	0.206†	CD36	GATA1	0.170†			
	CD41#	MYB	−0.129†	TUBB2A	RUNX1	−0.020			
	TUBB2A	RUNX1	0.109						
	CD44	ETS1	0.086†						
	TGFB1	DDB1	0.077†						
$\widehat{\Theta}$	GFI1B	RUNX1	0.274†	MYB	TAL1	0.370	FOG1#	GATA2	0.314†
	ETS1	TAL1	0.241	CDK6	GATA2	0.297†	CDK4	GATA2	0.245†
	MYB	TAL1	0.215	CDK4	P27KIP1#	0.225†	ETS1	TAL1	0.238
	DDB1	MYB	0.206†	CD34	GATA2	0.222	GFI1B	RUNX1	0.237†
	GATA2	TAL1	0.204	GATA1	TAL1	0.182†	SPI1	TAL1	0.175
	GATA2	MYB	0.201†	ETS1	TAL1	0.181	GATA1	TAL1	0.165†
	CDK6	GATA2	0.181†	GATA1	GATA2	0.179†	GATA1	PIM1	0.161†
	CTNNB1	MYB	0.169†	FOXO3	GATA2	0.145	GATA2	P27KIP1#	0.153
	LSD1#	SPI1	0.168†	SPI1	TAL1	0.141	DDB1	MYB	0.141†
	GATA2	GFI1B	0.164†	CD34	MYB	0.110	CD34	MYB	0.138
	CD34	MYB	0.148	GATA2	TAL1	0.082	GATA2	TAL1	0.128
	MEIS1	MYB	0.135†	RUNX1	TAL1	0.070	CDK6	MYB	0.126†
	GFI1B	LSD1#	0.131†	KLF1	TAL1	−0.056†	FOXO3	GATA2	0.125
	GATA2	SPI1	0.095	GATA2	P27KIP1#	0.053	RUNX1	TAL1	0.116
	FOXO3	GATA2	0.095	GATA2	SPI1	0.038	CDK4	P27KIP1#	0.114†
	GATA2	P27KIP1#	0.092	CDK6	TAL1	0.001†	MEIS1	MYB	0.107†
	CDK6	MYB	0.085†				GATA1	GATA2	0.089†
	RUNX1	TAL1	0.080				GATA2	SPI1	0.084
	MYB	SPI1	0.064†				GATA2	GFI1B	−0.077†
	JAK2	MYB	0.063†				JAK2	MYB	0.062†
	SPI1	TAL1	0.055				GFI1B	MYB	0.062†

(*continued*)

Table 3. Continued

Pre-MEP			E-MEP			MK-MEP		
from	to	PC	from	to	PC	from	to	PC
GFI1B	MYB	0.053†				MYB	SPI1	0.061†
GFI1B	SPI1	0.040†				CD34	GATA2	0.060
CD34	GATA2	0.026				FLI1	FOXO3	0.055†
CDK6	TAL1	−0.020†				MYB	TAL1	−0.049
CTNNB1	GATA2	0.010†				GATA2	MYB	0.045†
CDK4	GATA2	0.009†						
GATA2	LSD1#	0.004†						

Note. Each column corresponds to a sub-population (Pre-MEP, E-MEP, and MK-MEP); each row shows the undirected links among the nuclear ($\widehat{\Theta}$) and membrane ($\widehat{\Omega}$) factors, respectively, and the directed links between them ($\widehat{B}$). The 'PC' columns report the partial correlation coefficients for $\widehat{\Theta}$ and $\widehat{\Omega}$, and the regression coefficients for $\widehat{B}$. The dagger symbols at the top highlight specific links in each sub-population.
CD41/ITGA2B, CD42/GP1BA, CD61/ITGB3, CD71/TFRC, CD105/ENG, CD110/MPL, CD117/KIT, FOG1/ZFPM1, LSD1/KDM1A, P27KIP1/CDKN1B. Pre-MEP = pre-megakaryocyte-erythroid progenitor (MEP); E-MEP = erythroid-MEP; MK-MEP = megakaryocyte-MEP.

support the fact that pre-MEP is prone to differentiate towards the E-MEP lineage (Psaila et al., 2016), while further regulation is required to confirm commitment towards the MK-MEP lineage (Izzi et al., 2021).

9 Conclusion

Motivated by the study of blood cell formation, and the search for regulatory mechanisms that underlie the differentiation of progenitor stem cells into more mature blood cells, we have proposed a complex graphical model that allows to infer interactions that are specific to each cellular lineage, both within and between proteins belonging to different types, here nuclear factors and membrane receptors that are known to be involved in human haematopoiesis. We have devised a computationally efficient strategy for the inference of the network of dependences in a highly nonstandard setting, characterized by high dimensionality, both at the level of responses and covariates, the presence of distinct sub-populations of cells, and missingness, which is typical of RT-qPCR data. The algorithm is based on a careful combination of alternating direction method of multipliers algorithms, that allows to achieve sparsity in the lineage-specific networks within, and the associations between, the membrane-bound receptors and the nuclear factors as well as a high sharing between networks across the different sub-populations. The approach can be used on similar applications where network inference is conducted from high-dimensional, heterogeneous and partially observed data.

Acknowledgments

The authors are thankful to the joint editor, the associate editor, and the anonymous reviewers for their valuable comments.

Conflicts of interest: None declared.

Funding

L.A. and G.S. gratefully acknowledge financial support from the University of Palermo (FFR2022, FFR2023 and FFR2024). L.A. and G.S. were also financially supported by the European Union - Next Generation EU - Mission 4 Component 2 - CUP: B53D23009480006. V.V. was financially supported by the European Union - Next Generation EU - Mission 4 Component 2 - CUP: C53D23002580006. G.S. acknowledges support by the Italian Ministry of University and Research (MUR) through the project PON-AIM "Attraction and International Mobility":

AIM1873193-2 activity 1. C.C. and W.A. acknowledge support by Regione Siciliana, through the PO FESR action 1.1.5, project OBIND N.086202000366—CUP G29J18000700007.

Data availability

The computational approach presented in the paper will be implemented in the `R` package `cglasso`. The code for replicating the analysis presented in this paper is openly available at the following `GitHub` repository https://github.com/gianluca-sottile/Hematopoiesis-network-inference-from-RT-qPCR-data.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series C*.

References

Augugliaro L., Abbruzzo A., & Vinciotti V. (2020). ℓ_1-penalized censored gaussian graphical model. *Biostatistics*, *21*, e1–e16. https://doi.org/10.1093/biostatistics/kxy043

Augugliaro L., Sottile G., & Vinciotti V. (2020). The conditional censored graphical lasso estimator. *Statistics and Computing*, *30*(5), 1273–1289. https://doi.org/10.1007/s11222-020-09945-7

Augugliaro L., Sottile G., Wit E. C., & Vinciotti V. (2023). cglasso: An `R` package for conditional graphical lasso inference with censored and missing values. *Journal of Statistical Software*, *105*(1), 1–58. https://doi.org/10.18637/jss.v105.i01

Behrouzi P., & Wit E. C. (2019). Detecting epistatic selection with partially observed genotype data by using copula graphical models. *Journal of the Royal Statistical Society: Series C*, *68*(1), 141–160. https://doi.org/10.1111/rssc.12287

Blank U., & Karlsson S. (2015). TGF-β signaling in the control of hematopoietic stem cells. *Blood*, *125*(23), 3542–3550. https://doi.org/10.1182/blood-2014-12-618090

Boyd S., Parikh N., & Chu E. (2011). *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc.

Boyer T. C., Hanson T., & Singer R. S. (2013). Estimation of low quantity genes: A hierarchical model for analyzing censored quantitative real-time PCR data. *PloS One*, *8*(5), e64900. https://doi.org/10.1371/journal.pone.0064900

Chen M., Ren Z., Zhao H., & Zhou H. (2016). Asymptotically normal and efficient estimation of covariate-adjusted Gaussian graphical model. *Journal of the American Statistical Association*, *111*(513), 394–406. https://doi.org/10.1080/01621459.2015.1010039

Cheng H., Zheng Z., & Cheng T. (2020). New paradigms on hematopoietic stem cell differentiation. *Protein Cell*, *11*(1), 34–44. https://doi.org/10.1007/s13238-019-0633-0

Chiquet J., Huard T. M., & Robin S. (2017). Structured regularization for conditional Gaussian graphical models. *Statistics and Computing*, *27*(3), 789–804. https://doi.org/10.1007/s11222-016-9654-1

Chun H., Chen M., Li B., & Zhao H. (2013). Joint conditional Gaussian graphical models with multiple sources of genomic data. *Frontiers in Genetics*, *4*, 1–8. https://doi.org/10.3389/fgene.2013.00294

Danaher P., Wang P., & Witten D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B*, *76*(2), 373–397. https://doi.org/10.1111/rssb.12033

Doré L. C., & Crispino J. D. (2011). Transcription factor networks in erythroid cell and megakaryocyte development. *Blood*, *118*(2), 231–239. https://doi.org/10.1182/blood-2011-04-285981

Genz A., & Bretz F. (2002). Comparison of methods for the computation of multivariate *t* probabilities. *Journal of Computational and Graphical Statistics*, *11*(4), 950–971. https://doi.org/10.1198/106186002394

Guo J., Levina E., Michailidis G., & Zhu J. (2011). Joint estimation of multiple graphical models. *Biometrika*, *98*(1), 1–15. https://doi.org/10.1093/biomet/asq060

Guo J., Levina E., Michailidis G., & Zhu J. (2015). Graphical models for ordinal data. *Journal of Computational and Graphical Statistics*, *24*, 183–204. https://doi.org/10.1080/10618600.2014.889023

Huang F., Chen S., & Huang S. J. (2018). Joint estimation of multiple conditional Gaussian graphical models. *IEEE Transactions on Neural Networks and Learning Systems*, *29*, 3034–3046. https://doi.org/10.1109/TNNLS.2017.2710090

Ibrahim J. G., Zhu H., & Tang N. (2008). Model selection criteria for missing-data problems using the EM algorithm. *Journal of the American Statistical Association*, *103*, 1648–1658. https://doi.org/10.1198/016214508000001057

Izzi B., Gialluisi A., Gianfagna F., Orlandi S., De Curtis A., Magnacca S., Costanzo S., Di Castelnuovo A., Donati M. B., de Gaetano G., Hoylaerts M. F., Cerletti C., & Iacoviello L. (2021). Platelet distribution width is

associated with P-selectin dependent platelet function: Results from the Moli-family cohort study. *Cells*, *10*(10), 2737. https://doi.org/10.3390/cells10102737

Kling T., Johansson P., Sanchez J., Marinescu V. D., Jörnsten R., & Nelander S. (2015). Efficient exploration of pan-cancer networks by generalized covariance selection and interactive web content. *Nucleic Acids Research*, *43*(15), e98–e98. https://doi.org/10.1093/nar/gkv413

Krämer A., Green J., Pollard Jr J., & Tugendreich S. (2014). Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics*, *30*(4), 523–530. https://doi.org/10.1093/bioinformatics/btt703

Lafferty J., McCallum A., & Pereira F. C (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001)* (pp. 282–289).

Lauritzen S. L. (1996). *Graphical models*. Oxford University Press.

Lee W., & Liu Y. (2015). Joint estimation of multiple precision matrices with common structures. *Journal of Machine Learning Research*, *16*, 1035–1062.

Li B., Chun H., & Zhao H. (2012). Sparse estimation of conditional graphical models with application to gene networks. *Journal of the American Statistical Association*, *107*(497), 152–167. https://doi.org/10.1080/01621459.2011.644498

Lin J., Basu S., Banerjee M., & Michailidis G. (2016). Penalized maximum likelihood estimation of multi-layered gaussian graphical models. *Journal of Machine Learning Research*, *17*, 1–51.

Little R. J. (2021). Missing data assumptions. *Annual Review of Statistics and Its Application*, *8*(1), 89–107. https://doi.org/10.1146/statistics.2021.8.issue-1

Little R. J. A., & Rubin D. B. (2002). *Statistical analysis with missing data* (2nd ed.). John Wiley & Sons, Inc.

Ma J., & Michailidis G. (2016). Joint structural estimation of multiple graphical models. *Journal of Machine Learning Research*, *17*, 1–48.

Majumdar S., & Michailidis G. (2022). Joint estimation and inference for data integration problems based on multiple multi-layered gaussian graphical models. *Journal of Machine Learning Research*, *23*, 1–53.

Matsumoto A., & Nakayama K. I. (2013). Role of key regulators of the cell cycle in maintenance of hematopoietic stem cells. *Biochimica et Biophysica Acta (BBA)-General Subjects*, *1830*(2), 2335–2344. https://doi.org/10.1016/j.bbagen.2012.07.004

McCall M. N., McMurray H. R., Land H., & Almudevar A. (2014). On non-detects in qPCR data. *Bioinformatics*, *30*(16), 2310–2316. https://doi.org/10.1093/bioinformatics/btu239

McLachlan G., & Krishnan T. (2008). *The EM algorithm and extensions* (2nd ed.). John Wiley & Sons, Inc.

Meng Y., Pospiech M., Ali A., Chandwani R., Vergel M., Onyemaechi S., Yaghmour G., Lu R., & Alachkar H. (2023). Deletion of cd36 exhibits limited impact on normal hematopoiesis and the leukemia microenvironment. *Cellular & Molecular Biology Letters*, *28*, 1–21. https://doi.org/10.1186/s11658-023-00455-8

Mohammadi R., & Wit E. (2019). BDgraph: An R package for Bayesian structure learning in graphical models. *Journal of Statistical Software*, *89*, 1–30. https://www.jstatsoft.org/article/view/v089i03. https://doi.org/10.18637/jss.v089.i03

Psaila B., Barkas N., Iskander D., Roy A., Anderson S., Ashley N., Caputo V. S., Lichtenberg J., Loaiza S., Bodine D. M., Karadimitris A., Mead A. J., & Roberts I. (2016). Single-cell profiling of human megakaryocyte-erythroid progenitors identifies distinct megakaryocyte and erythroid differentiation pathways. *Genome Biology*, *17*, 1–19. https://doi.org/10.1186/s13059-016-0939-7

Rothman A. J., Levina E., & Zhu J. (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, *19*, 947–962. https://doi.org/10.1198/jcgs.2010.09188

Samanta S., Khare K., & Michailidis G. (2022). A generalized likelihood-based Bayesian approach for scalable joint regression and covariance selection in high dimensions. *Statistics and Computing*, *32*, 47. https://doi.org/10.1007/s11222-022-10102-5

Scheicher R., Hoelbl-Kovacic A., Bellutti F., Tigan A.-S., Prchal-Murphy M., Heller G., Schneckenleithner C., Salazar-Roa M., Zöchbauer-Müller S., Zuber J., Malumbres M., Kollmann K., & Sexl V. (2015). CDK6 as a key regulator of hematopoietic and leukemic stem cell activation. *Blood*, *125*, 90–101. https://doi.org/10.1182/blood-2014-06-584417

Sherina V., McMurray H. R., Powers W., Land H., Love T. M., & McCall M. N. (2020). Multiple imputation and direct estimation for qPCR data with non-detects. *BMC Bioinformatics*, *21*, 1–15. https://doi.org/10.1186/s12859-020-03807-9

Shin J. Y., Hu W., Naramura M., & Park C. Y. (2014). High c-Kit expression identifies hematopoietic stem cells with impaired self-renewal and megakaryocytic bias. *Journal of Experimental Medicine*, *211*, 217–231. https://doi.org/10.1084/jem.20131128

Sohn K.-A., & Kim S (2012). Joint estimation of structured sparsity and output structure in multiple-output regression via inverse-covariance regularization. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics* (eds. N. D. Lawrence and M. Girolami), vol. 22 of *Proceedings of Machine Learning Research* (pp. 1081–1089). PMLR.

Städler N., & Bühlmann P. (2012). Missing values: Sparse inverse covariance estimation and an extension to sparse regression. *Statistics and Computing*, *22*, 219–235. https://doi.org/10.1007/s11222-010-9219-7

Stekhoven D. J., & Buehlmann P. (2012). Missforest - non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*, 112–118. https://doi.org/10.1093/bioinformatics/btr597

Vagapova E., Spirin P., Lebedev T., & Prassolov V. (2018). The role of TAL1 in hematopoiesis and leukemogenesis. *Acta Naturae*, *10*, 15–23. https://doi.org/10.32607/20758251-2018-10-1-15-23

Wang J. (2015). Joint estimation of sparse multivariate regression and conditional graphical models. *Statistica Sinica*, *25*, 831–851. https://dx.doi.org/10.5705/ss.2013.192

Xie Y., Liu Y., & Valdar W. (2016). Joint estimation of multiple dependent gaussian graphical models with applications to mouse genomics. *Biometrika*, *103*, 493–511. https://doi.org/10.1093/biomet/asw035

Yin J., & Li H. (2011). A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *The Annals of Applied Statistics*, *5*, 2630–2650. https://doi.org/10.1214/11-AOAS494

Yin J., & Li H. (2013). Adjusting for high-dimensional covariates in sparse precision matrix estimation by ℓ_1-penalization. *Journal of Multivariate Analysis*, *116*, 365–381. https://doi.org/10.1016/j.jmva.2013.01.005

Yuan M., & Lin Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, *94*(1), 19–35. https://doi.org/10.1093/biomet/asm018

Zhang Y., Ouyang Z., & Zhao H. (2017). A statistical framework for data integration through graphical models with application to cancer genomics. *The Annals of Applied Statistics*, *11*, 161. https://doi.org/10.1214/16-AOAS998

Zhu Y., Shen X., & Pan W. (2014). Structural pursuit over multiple undirected graphs. *Journal of the American Statistical Association*, *109*(508), 1683–1696. https://doi.org/10.1080/01621459.2014.921182