Published in final edited form as:

Stat Med. 2020 January 30; 39(2): 146–155. doi:10.1002/sim.8408.

Integrating approximate single factor graphical models

Xinyan Fan¹, Kuangnan Fang^{2,3}, Shuangge Ma⁴, Qingzhao Zhang^{2,3,5}

¹School of Statistics, Renmin University of China, Beijing, China

²Department of Statistics, School of Economics, Xiamen University, Xiamen, China

³Key Laboratory of Econometrics, Ministry of Education, Xiamen University, Xiamen, China

⁴Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut

⁵The Wang Yanan Institute for Studies in Economics, Xiamen University, Xiamen, China

Abstract

In the analysis of complex and high-dimensional data, graphical models have been commonly adopted to describe associations among variables. When common factors exist which make the associations dense, the single factor graphical model has been proposed, which first extracts the common factor and then conducts graphical modeling. Under other simpler contexts, it has been recognized that results generated from analyzing a single dataset are often unsatisfactory, and integrating multiple datasets can effectively improve variable selection and estimation. In graphical modeling, the increased number of parameters makes the “lack of information” problem more severe. In this article, we integrate multiple datasets and conduct the approximate single factor graphical model analysis. A novel penalization approach is developed for the identification and estimation of important loadings and edges. An effective computational algorithm is developed. A wide spectrum of simulations and the analysis of breast cancer gene expression datasets demonstrate the competitive performance of the proposed approach. Overall, this study provides an effective new venue for taking advantage of multiple datasets and improving graphical model analysis.

Keywords

approximate single factor graphical model; integrative analysis; penalized high dimensional analysis

Correspondence Qingzhao Zhang, Department of Statistics, School of Economics, Xiamen University, Xiamen, China. qzzhang@xmu.edu.cn.

DATA AVAILABILITY STATEMENT

The analyzed data are publicly available from GEO(Gene Expression Omnibus). Any interested researcher will be able to access. However, we do not have authority to re-distribute them.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

1 | INTRODUCTION

In the analysis of complex and high-dimensional data, an important step is to understand how variables are associated with each other. Such analysis not only can lead to a better understanding of the underlying data generating mechanisms but also serve as the building block of downstream analysis, such as clustering and regression analysis. For describing the interconnections among variables, the graph (network) approach has been commonly adopted. In graph construction, there are two families of approaches. The first family conducts unconditional construction. That is, when examining the association between two variables, other variables are ignored. The second family conducts conditional construction and examines whether two variables are conditionally connected after adjusting for the effects of other variables. The conditional construction can be more informative and, at the same time, methodologically and numerically more challenging. We refer to the work of Fan et al¹ for relevant discussions.

Among the available conditional constructions, a popular approach is the Gaussian graphical model (GGM). Under the “classic” GGM, variables are assumed to have a joint normal distribution. Here, two variables are conditionally independent (and hence, not connected in the graph) if the corresponding element in the precision matrix, which is the inverse of the covariance matrix, is zero. As such, constructing the graph structure amounts to a sparse estimation of the precision matrix. Extensive research has been conducted on the methodology,² computation,^{3,4} and theory^{5,6} of the GGM. In addition, methodologies have also been developed to accommodate data that are not normally distributed.⁷

In practical data analysis, it has been found that variables sometimes share common factors that are “less interesting.” In, for example, the analysis of genetic data, such common factors can be related to the batch effects in profiling or the same (or related) genetic ancestry shared by samples. The presence of such common factors may make variables artificially densely connected and mask truly interesting connections. Multiple studies have suggested that such common factors should be removed prior to constructing graphs. To this end, the approximate factor graphical model has been proposed. This approach assumes that the underlying data generation mechanism can be decomposed into two parts: a shared common factor and a sparse graph, both of which are unknown and need to be jointly estimated. More details can be found in Section 2. We also refer to the work of Fan et al⁸ for developments on the approximate factor graphical model.

Estimation with high-dimensional data in general suffers from a “lack of information.” This problem gets more severe with graphical models. Specifically, with p variables, a regression analysis involves $O(p)$ unknown parameters, whereas a graphical model involves $O(p^2)$ parameters. When the sample size n is small to moderate, it has been recognized that results generated from a single dataset can be unsatisfactory. For regression, clustering, and other problems, integrative analysis, under which raw data from multiple independent studies with comparable designs are jointly analyzed, has emerged as a powerful tool and has been shown to outperform single-dataset analysis and other multi-datasets analysis including meta-analysis. For discussions on integrative analysis under regression and other “simpler” settings, we refer to the works of Liu et al,⁹ Zhao et al,¹⁰ and Fang et al.¹¹

With graphical models, integrative analysis has also been conducted.^{12,13} However, it is noted that such analysis has not been well conducted with the approximate single factor graphical model, which has one more “layer” of complexity. In addition, carefully examining the literature suggests that, in the existing integrative analysis with graphical models, there has been insufficient attention to the “interconnections” among datasets. Specifically, with the independence of the analyzed datasets, the existing studies treat parameters in different datasets as “unrelated.” It has been well noted in the literature that the similarity among datasets is the basis of integrative and other multidatasets analysis. Such as, under certain scenarios, it is sensible to consider the “interconnections” among parameters in different datasets in estimation. Similar considerations have been taken in simple regression settings¹⁴ and led to improved estimation and variable selection.

In this article, we conduct the integrative analysis of multiple datasets under the approximate single factor graphical models. This study will extend the powerful integrative analysis paradigm to the graphical models and provide a useful new way for studying the interconnections among variables. It will advance from the existing integrative analysis of regression models by considering more complex graphical models and, from the existing integrative analysis of graphical models, by accounting for the shared common factors and, more importantly, the “interconnections” among parameters in different datasets. Extensive methodological and numerical studies will be conducted to establish the competitive performance of the proposed approach. Overall, this study may provide a useful new venue for using graphical tools to study variable connections.

2 | METHODS

Denote K as the number of independent datasets. For simplicity of notation, assume that the same set of p variables is measured in all datasets. Here, we note that integrative analysis may involve multiple nontrivial practical issues, such as the selection of comparable datasets, matching variables across datasets, and accommodating unmatched variables. We fully acknowledge the challenge of these issues. However, since they have been discussed in detail in the literature,¹⁵ we choose to avoid overly redundant discussions.

For dataset $k (= 1, \dots, K)$, consider the approximate single factor model

$$\mathbf{y}^{(k)} = \mathbf{b}^{(k)} f^{(k)} + \epsilon^{(k)}, \quad k = 1, \dots, K, \quad (1)$$

where $\mathbf{y}^{(k)} = (y_1^{(k)}, \dots, y_p^{(k)})^T$ is the p -vector of observations (“response variables”), $\mathbf{b}^{(k)} = (b_1^{(k)}, \dots, b_p^{(k)})^T$ is the vector of factor loadings, and $f^{(k)}$ is the common factor with mean 0 and variance 1. Overall, $\mathbf{b}^{(k)} f^{(k)}$ represents the shared common factors. $\epsilon^{(k)}$ is the idiosyncratic component with mean $\mathbf{0}$ and covariance matrix $\Sigma^{(k)}$, uncorrelated with $f^{(k)}$. Certain assumptions would be needed prior to estimation. For example, to ensure the identifiability of the common and idiosyncratic components, it needs to be assumed that $\|\mathbf{b}^{(k)}\|$, the ℓ_2 norm of $\mathbf{b}^{(k)}$, is much larger than the maximal eigenvalue of $\Sigma^{(k)}$. It is noted that such assumptions are “inherent” to the approximate single factor graphical models, as opposed to integrative analysis.

Denote $\Theta^{(k)} = (\Sigma^{(k)})^{-1}$. It describes the interconnections among variables in dataset k after removing the effect of the shared factors. In our analysis, similar to that for a single dataset, our goal is to estimate $\mathbf{b}^{(k)}$ and $\Theta^{(k)}$. Following the works of Fan et al⁸ and Hirose and Yamamoto,¹⁶ we consider the case where both $\mathbf{b}^{(k)}$ and $\Theta^{(k)}$ are sparse.

Consider the objective function

$$\sum_{k=1}^K \left\{ -n_k \log \det(\Theta^{(k)}) + \sum_{i=1}^{n_k} \left[\left(\mathbf{y}_i^{(k)} - \mathbf{b}^{(k)} f_i^{(k)} \right)^T \Theta^{(k)} \left(\mathbf{y}_i^{(k)} - \mathbf{b}^{(k)} f_i^{(k)} \right) \right] \right\} + P(\{\mathbf{b}\}, \{\Theta\}; \lambda), \quad (2)$$

subject to the constraints that $\Theta^{(k)}$ is positive definite and $1 / n_k \sum_{i=1}^{n_k} f_i^{(k)2} = 1$ for all k . In (2), n_k is the sample size of the k th dataset, $\mathbf{y}_i^{(k)} = (y_{i1}^{(k)}, \dots, y_{ip}^{(k)})^T$ is the i th observation of $\mathbf{y}^{(k)}$, $f_i^{(k)}$ is the i th factor score, $\{\mathbf{b}\} = \{\mathbf{b}^{(k)}, k = 1, \dots, K\}$, $\{\Theta\} = \{\Theta^{(k)}, k = 1, \dots, K\}$, and λ is the (vector of) tuning associated with penalty function $P(\cdot)$. In (2), the first term is the sum of K loss functions, with one for each dataset. The development of this loss function has been examined in detail in the works of Rothman et al¹⁷ and Yin and Li,¹⁸ and will not be reiterated here. As in quite a few published studies,¹⁹ we adopt the penalization technique for regularized estimation and selection of nonzero effects. As has been noted in the literature, the key challenge lies in properly designing the penalty function.

We first consider the penalty

$$4\lambda_1 \sum_{j=1}^p \left(\sum_{k=1}^K |b_j^{(k)}| \right)^{1/2} + 4\lambda_2 \sum_{j \neq \ell} \left(\sum_{k=1}^K |\theta_{j\ell}^{(k)}| \right)^{1/2}, \quad (3)$$

where $b_j^{(k)}$ is the j th component of $\mathbf{b}^{(k)}$, $\theta_{j\ell}^{(k)}$ is the (j, ℓ) th element of $\Theta^{(k)}$, and $\lambda_1 > 0$ and $\lambda_2 > 0$ are data-dependent tuning parameters. Here, the group bridge penalty is adopted. For a specific element of the loading (and a specific element of the precision matrix), we treat its K values in the K datasets as a group. The group-level penalization determines whether this element is important at all. Furthermore, the within-group-level penalization determines, for an important element, in which dataset(s) it is nonzero. With this two-level selection, this penalization approach intrinsically assumes the heterogeneity structure, that is, multiple datasets can have overlapping but different sparsity structures. We refer to the works of Ma et al²⁰ and Guo et al¹² for applications of the group bridge to integrative analysis under simpler data/model settings.

With the group bridge penalty, the “interconnections” among datasets are considered via the grouping structures, which may be insufficient. For integrative analysis as well as other multidatasets analysis, the similarity across datasets is the foundation of analysis. Under certain scenarios, for example, when the analysis of meta data suggests a high level of similarity in study design and sample characteristics, it can be sensible to expect that the loading vectors and precision matrices in different datasets are similar in magnitude. In this case, we propose further adding the following penalty to (3):

$$\lambda_3 \left[\sum_{j=1}^p \sum_{k < k'} (b_j^{(k)} - b_j^{(k')})^2 + \sum_{j \neq \ell} \sum_{k < k'} (\theta_{j\ell}^{(k)} - \theta_{j\ell}^{(k')})^2 \right], \quad (4)$$

where $\lambda_3 > 0$ is a data-dependent tuning parameter. The proposed penalty takes a fused form and directly encourages similar magnitudes across datasets. In the literature, there are other types of fused penalties, for example, those based on ℓ_1 norm. In our analysis, our goal is to achieve similarity in magnitude, and exact equality, which can be achieved by ℓ_2 type penalties, is of less interest. It is also noted that the ℓ_2 form of penalty may also facilitate computation.

Remark 1. The proposed approach involves three penalties. With the complexity of the approximate single factor graphical model as well as the need to accommodate across-dataset interconnections, it is inevitable that the approach gets more complicated. Recent published studies have shown that, methodologically and computationally, it is manageable to have three penalties.²¹⁻²³ To reduce computational cost, it is also possible to set λ_1 and λ_2 as connected. In this paper, we assume that a single factor (per dataset) is sufficient to remove the shared common effects. The proposed approach (including the computational algorithm described below) can be extended to multifactor models.

Computation.

Denote $Y^{(k)} = (\mathbf{y}_1^{(k)}, \dots, \mathbf{y}_{n_k}^{(k)})$, $\mathbf{f}^{(k)} = (f_1^{(k)}, \dots, f_{n_k}^{(k)})^\top$ and $\{\mathbf{f}\} = \{\mathbf{f}^{(k)}, k = 1, \dots, K\}$. The proposed algorithm is summarized in Algorithm 1. The details for Step 2(a) and 2(b) are provided in Supplementary Materials. Overall, we take an iterated strategy, optimize with respect to one set of parameters at a time, while keeping the other set as fixed.

Algorithm 1

1. Initialize: $m = 0$, $\Theta_{(m)} = \mathbf{I}$, $\frac{1}{\sqrt{n_k}} \mathbf{f}^{(k)}$ as the first principal component of $Y^{(k)T} Y^{(k)}$ and $\mathbf{b}_{(m)} = 1 / n_k Y^{(k)} \mathbf{f}_{(m)}^{(k)}$ for $k = 1, \dots, K$.
 2. Update $m = m + 1$
 - (a) Update $\{\mathbf{b}\}_{(m)}$ and $\{\mathbf{f}\}_{(m)}$ as the minimizer of

$$\sum_{k=1}^K \sum_{i=1}^{n_k} \left[\left(\mathbf{y}_i^{(k)} - \mathbf{b}_{(m)} f_i^{(k)} \right)^\top \Theta_{(m-1)} \left(\mathbf{y}_i^{(k)} - \mathbf{b}_{(m)} f_i^{(k)} \right) + 4\lambda_1 \sum_{j=1}^p \left(\sum_{k=1}^K |b_j^{(k)}| \right)^{1/2} + \lambda_3 \sum_{j=1}^p \sum_{k < k'} (b_j^{(k)} - b_j^{(k')})^2 \right], \quad (5)$$
 subject to $\frac{1}{n_k} \sum_{i=1}^{n_k} f_i^{(k)2} = 1$, for $k = 1, \dots, K$.
 - (b) Calculate $S_{(m)}^{(k)} = 1 / n_k \sum_{i=1}^{n_k} (\mathbf{y}_i^{(k)} - \mathbf{b}_{(m)} f_i^{(k)}) (\mathbf{y}_i^{(k)} - \mathbf{b}_{(m)} f_i^{(k)})^\top$. Update $\{\Theta\}_{(m)}$ as the minimizer of

$$\min \sum_{k=1}^K -n_k \left\{ \log \det(\Theta^{(k)}) - \text{tr} \left(S_{(m)}^{(k)} \Theta^{(k)} \right) \right\} + 4\lambda_2 \sum_{j \neq \ell} \left(\sum_{k=1}^K |\theta_{j\ell}^{(k)}| \right)^{1/2} + \lambda_3 \sum_{j \neq \ell} \sum_{k < k'} (\theta_{j\ell}^{(k)} - \theta_{j\ell}^{(k')})^2. \quad (6)$$
 3. Repeat Step 2 until convergence, which is concluded when the difference between two consecutive estimates is smaller than a pre-defined threshold.
-

For the overall algorithm, the objective function is bounded below, and its value decreases at each iteration. For the updates in Steps 2(a) and 2(b), existing studies have shown that the technique has satisfactory convergence properties. As such, the proposed algorithm is

expected to have satisfactory properties. Examining Algorithm 1 and algorithms for Steps 2(a) and 2(b) suggests that only simple updates are involved. The overall computational cost is thus affordable. For the analysis of one simulated replicate with three datasets (more details in Section 3), the proposed analysis can be finished within 6.28 minutes on a regular desktop.

The proposed approach involves tuning parameters, controlling selection, regularized estimation, and similarity across datasets. In our numerical study, we experiment with multiple commonly adopted criteria and find that Akaike information criterion (AIC) leads to the best performance. The tuning parameters selected are stable in the three-dimensional parametric space. We adopt AIC in our simulation and data analysis but recommend that other criteria also should be considered in practical data analysis for cautions.

3 | SIMULATION

We conduct simulation to assess performance of the proposed approach and compare with alternatives. We set $K = 3$, $p = 100$, and $n_k = n = 200$ for $k = 1, \dots, K$. $\mathbf{y}^{(k)}$'s are randomly drawn from $\mathcal{N}(0, \mathbf{b}^{(k)}\mathbf{b}^{(k)\top} + \Sigma^{(k)})$. Specifically, we set $\mathbf{b}^{(k)} = 5\mathbf{b}^{*(k)} / \|\mathbf{b}^{*(k)}\|$ and consider the following cases: (a) $b_i^{*(1)} = b_i^{*(2)} = b_i^{*(3)} \sim \mathcal{N}(1, 1/3)$ for $i = 1, \dots, M$ and $b_i^{*(k)} = 0$ for $i = (M+1), \dots, p$, $k = 1, \dots, K$ with $M = p/4, p/2$, and $3p/4$; and (b) $b_i^{*(1)} = b_i^{*(2)} = b_i^{*(3)} \sim \mathcal{N}(1, 1/3)$ for $i = (p/8 + 1), \dots, (p/2)$, $b_i^{*(1)} \sim \mathcal{N}(1, 1/3)$ for $i = 1, \dots, (p/8)$, $b_i^{*(3)} \sim \mathcal{N}(1, 1/3)$ for $i = (p/2 + 1), \dots, (5p/8)$, and otherwise $b_i^{*(k)} = 0$ for $k = 1, \dots, K$. Under case (a), $\mathbf{b}^{(k)}$'s have the same sparsity structure as well as the same values across K datasets and the sparsity degree is controlled by M . Under case (b), $\mathbf{b}^{(k)}$'s have partially overlapping sparsity structures and similar (but not the same) values for the overlapping nonzero elements.

The scale-free and Erdos-Renyi graph structures, which are among the most popular, are considered for $\Theta^{(1)} = (\Sigma^{(1)})^{-1}$. There are extensive discussions on these two structures in the literature. Very briefly, the scale-free graph is generated with one edge added in each step, whereas the Erdos-Renyi graph is generated with a probability of 0.05 for drawing an edge between any two graph nodes. For a given graph structure, the corresponding precision matrix is generated as follows. The $p \times p$ matrix $\Theta^{(1)}$ is first created. The elements not corresponding to edges are set as zero. For elements corresponding to edges, values are generated randomly from $\mathcal{U}([-0.4, -0.1] \cup [0.1, 0.4])$. Further, heterogeneity is added to the structure for the first dataset to generate $\Theta^{(2)}$ and $\Theta^{(3)}$. Specifically, for $\Theta^{(2)}$ and $\Theta^{(3)}$, a pair of symmetric zero elements in $\Theta^{(1)}$ are randomly selected and replaced with a value randomly generated from $\mathcal{U}([-0.4, -0.1] \cup [0.1, 0.4])$. This procedure is repeated $p_d s_0$ times, where s_0 is the number of edges in $\Theta^{(1)}$. For p_d , we consider $p_d = 0, 0.1, 0.3$, and 0.5 , which gradually increase the differences across datasets. To ensure positive-definiteness, we set $\theta_{jj}^{(k)} = |\min_k \phi(\Theta^{(k)})_{\min}| + 0.5$ for $j = 1, \dots, p$, $k = 1, \dots, K$, where $\phi(\Theta^{(k)})_{\min}$ is the smallest eigenvalue of $\Theta^{(k)}$. Finally, we compute $\Sigma^{(k)} = (\Theta^{(k)})^{-1}$.

When assessing performance of the proposed approach, we are first interested in the identification accuracy of graph structures. To get a comprehensive picture, the receiver operating characteristic (ROC) curve technique is used. In the proposed approach, the penalty corresponding to λ_3 is newly proposed. To more comprehensively appreciate its impact, we consider multiple fixed values. The values of λ_1 and λ_2 are then varied to generate various true positive rate (TPR) and false positive rate (FPR) values. Here, the TPR and FPR are computed as the averages across three datasets. To assess estimation performance, we consider estimation error (ER), which is defined as

$$\text{ER} = \frac{1}{K} \sum_{k=1}^K \sum_{i \neq j} \left(\theta_{ij}^{(k)} - \hat{\theta}_{ij}^{(k)} \right)^2 / \sum_{i \neq j} \left(\theta_{ij}^{(k)} \right)^2,$$

where $\theta_{ij}^{(k)}$ s and $\hat{\theta}_{ij}^{(k)}$ s are the (i, j) th elements of $\Theta^{(k)}$ and $\hat{\Theta}^{(k)}$, respectively. For ER, following the same spirit as with the ROC curve, we also consider its value as a function of FPR. In the second set of evaluation, we still consider a sequence of λ_3 values but select λ_1 and λ_2 using the AIC criterion. For $\{\mathbf{b}\}$, we consider the following measures: angle (which is the mean acute angle between the estimated and true $\mathbf{b}^{(k)}$ s), NR (the mean ratio of $\|\hat{\mathbf{b}}^{(k)}\| / \|\mathbf{b}^{(k)}\|$ s), TPR, and FPR. In addition, we also consider the ER, TPR, FPR, and Matthews correlation coefficient (MCC) score for estimating the graph structures. Here, MCC is defined as follows:

$$\text{MCC} = \frac{1}{K} \sum_{k=1}^K \frac{(\text{TP}^{(k)} * \text{TN}^{(k)} - \text{FP}^{(k)} * \text{FN}^{(k)})}{\sqrt{(\text{TP}^{(k)} + \text{FP}^{(k)}) * (\text{TP}^{(k)} + \text{FN}^{(k)}) * (\text{TN}^{(k)} + \text{FP}^{(k)}) * (\text{TN}^{(k)} + \text{FN}^{(k)})}},$$

where $\text{TP}^{(k)}$, $\text{TN}^{(k)}$, $\text{FP}^{(k)}$, and $\text{FN}^{(k)}$ are the numbers of true positives, true negatives, false positive, and false negative for dataset k . Multiple approaches are potentially applicable to the simulated data. For example, it is possible to analyze each dataset separately and then conduct meta-analysis. However, the superiority of integrative analysis over meta-analysis and some other multidatasets methods has been well established, and hence, such comparisons will not be conducted. In the literature, the most relevant competitor is the JMG approach in the work Guo et al.¹² The JMG jointly estimates the graphical models of multiple datasets, with the goal of preserving the common structure of the precision matrices. However, it does not consider the common factors. It is also noted that the proposed approach with $\lambda_3 = 0$ can be viewed as another alternative, which conducts integrative analysis with two-level selection but does not promote similarity in parameter magnitude.

The results are presented in Figure 1, Table 1, and figures and tables in the Supplementary Materials. Different scenarios have different numerical results. However, the overall patterns are similar, with the proposed approach outperforming JMG as well as the group bridge approach with $\lambda_3 = 0$. Consider, for example, Figure 1. The proposed approach has the ROC curves on the top while having the ER curves on the bottom. Different values of λ_3 lead to moderately different results. In Figure 1, a stronger shrinkage (a large λ_3 value) leads to better results, which is attributable to the similarity across datasets. When the similarity

decreases, a smaller λ_3 value may be favored. Superiority is also observed when the tuning parameters are selected using AIC. Consider, for example, Table 1 with $p_d = 0.3$. For the estimation of $\{\Theta\}$, JMG has ER 1.31, the group bridge approach has ER 0.92, whereas the proposed approach has ERs 0.74, 0.40, and 0.35 for different λ_3 values. For identification, JMG has MCC 0.26, the group bridge approach has MCC 0.33, compared to 0.35, 0.44, and 0.54 for the proposed approach. We have also experimented with a few other settings and made similar observations (details omitted).

4 | DATA ANALYSIS

To test its practical applicability, we apply the proposed approach to three breast cancer datasets from Gene Expression Omnibus (GEO), which is a National Center for Biotechnology Information database for gene expression data. The three datasets have GEO IDs GSE5364, GSE22820, and GSE15852. We refer to published studies²⁴⁻²⁶ and the GEO website for more information on data. Briefly, the datasets have sample size 196, 186, and 86, respectively, and have been jointly analyzed in published studies.²⁷ For each dataset, missing data are imputed. Genes are matched across datasets using Entrez ID (based on the NCBI Entrez database). A total of 12 429 genes are measured in all three datasets. Although it is in principle possible to analyze all genes, it has been suggested that estimating such a huge number of parameters based on such a small sample size may be unreliable. As such, we conduct a supervised prescreening using the sparse PCA technique²⁸ and identify 31 genes in each dataset. Combining the selection results across datasets, we have 90 genes for downstream analysis.

We apply the proposed approach as well as JMG. It is noted that, unlike in simulation, λ_3 is also selected using AIC along with λ_1 and λ_2 . The analysis results are summarized in Table 2. It is seen that different approaches lead to different findings. For estimating the common factors, the proposed approach generates relatively consistent findings across datasets. Specifically, the angles between pairs of factor loading vectors are 21.6 (GSE5364 and GSE22820), 19.1 (GSE5364 and GSE15852), and 23.1 (GSE22820 and GSE15852), respectively. The graphs constructed using JMG and the proposed approach are presented in Figure 2. The differences between graphs are presented in the Supplementary Materials. The proposed approach identifies sparser networks. It is reasonable to conjecture that this is because the proposed approach effectively removes edges caused by the shared common factors.

With practical data, it is difficult to objectively evaluate performance of different approaches. We consider the following resampling based evaluation, which may provide some information. Specifically, we split each dataset into a training and a testing subset, with sizes 4:1. The proposed approach and JMG are applied to the training data. This process is repeated 100 times. The mean (sd) values of the average numbers of edges across the three datasets of the JMG and proposed approach are 1014.2 (136.0) and 864.4 (53.1), respectively. It is observed that the proposed approach identifies significantly fewer edges. We then use the training data estimates and testing data to compute the negative log-likelihood statistics, which are defined as

$$\frac{1}{\sum_{k=1}^K n_{tk}} \sum_{k=1}^K n_{tk} [-\log \det(\hat{\Theta}^{(k)}) + \text{tr}((S_t^{(k)} - \hat{\mathbf{b}}^{(k)} \hat{\mathbf{b}}^{(k)\top}) \hat{\Theta}^{(k)})] \text{ and}$$

$$\frac{1}{\sum_{k=1}^K n_{tk}} \sum_{k=1}^K n_{tk} [-\log \det(\hat{\Theta}^{(k)}) + \text{tr}(S_t^{(k)} \hat{\Theta}^{(k)})] \text{ for the proposed method and JMG,}$$

respectively. Here, n_{tk} and $S_t^{(k)}$ are the testing data sample size and sample covariance matrix for $k = 1, \dots, K$. The JMG and proposed approach have mean (sd) values of this statistic as 148.4 (89.7) and 127.4 (97.1). The proposed approach is observed to have better prediction. Overall, the fewer numbers of edges may suggest that the proposed approach can effectively remove common factors and reduce the number of edges, and the improved prediction may provide additional support to the validity of analysis.

5 | CONCLUSION

In this article, we have conducted the integrative analysis of multiple datasets under the approximate single factor graphical models. It is noted that although integrative analysis has been conducted under simpler contexts and the approximate single factor graphical model has been considered for a single dataset, their “marriage” has not been well conducted in the literature. In addition, this study also significantly advances from the existing joint graphical models by considering the “interconnections” among the regression coefficients of multiple datasets. The proposed approach has an intuitive formulation, and an effective algorithm has been designed to make it practically feasible. A wide spectrum of simulations and data analysis have demonstrated its competitive performance.

This study can be potentially extended in multiple directions. Although the approximate single factor graphical model has demonstrated competitive practical performance, it is still worthwhile to investigate integrative analysis with other graph construction methods. In the proposed approach, different datasets are shrunk to each other in the same manner. The “distances” between datasets, which, for, example can be partly inferred from meta-data, are not all equal. It may be of interest to develop more data-adaptive shrinkage methods. In this article, we have focused on methodological and numerical developments. It may be of interest to investigate theoretical properties of the proposed approach.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

We thank the associate editor and reviewers for their careful review and insightful comments, which have led to a significant improvement of this article. The authors gratefully acknowledge the National Natural Science Foundation of China (11971404 and 71471152), the Humanity and Social Science Youth Foundation of Ministry of Education of China (19YJC910010), the Fundamental Research Funds for the Central Universities (20720171064, 20720171095, and 20720181003), and the National Institutes of Health (CA216017).

Funding information

National Natural Science Foundation of China, Grant/Award Number: 11971404 and 71471152; Humanity and Social Science Youth Foundation of Ministry of Education of China, Grant/Award Number: 19YJC910010; Fundamental Research Funds for the Central Universities, Grant/Award Number: 20720171064, 20720171095, and 20720181003; National Institutes of Health, Grant/Award Number: CA216017

REFERENCES

1. Fan J, Liao Y, Liu H. An overview of the estimation of large covariance and precision matrices. *Econom J*. 2016;19(1):C1–C32.
2. Yuan M, Lin Y. Model selection and estimation in the gaussian graphical model. *Biometrika*. 2007;94(1):19–35.
3. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*. 2008;9(3):432–441. [PubMed: 18079126]
4. Witten D, Friedman J, Simon N. New insights and faster computations for the graphical lasso. *J Comput Graph Stat*. 2011;20(4):892–900.
5. Rothman AJ, Bickel PJ, Levina E, Zhu J. Sparse permutation invariant covariance estimation. *Electron J Stat*. 2008;2:494–515.
6. Lam C, Fan J. Sparsistency and rates of convergence in large covariance matrix estimation. *Ann Stat*. 2009;37(6B):4254–4278. [PubMed: 21132082]
7. Xue L, Zou H. Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Ann Stat*. 2012;40(5):2541–2571.
8. Fan J, Liu H, Wang W. Large covariance estimation through elliptical factor models. *Ann Stat*. 2018;46(4):1383–1414. [PubMed: 30214095]
9. Liu J, Ma S, Huang J. Integrative analysis of cancer diagnosis studies with composite penalization. *Scand Stat Theory Appl*. 2014;41(1):87–103. [PubMed: 24578589]
10. Zhao Q, Shi X, Huang J, Liu J, Li Y, Ma S. Integrative analysis of ‘-omics’ data using penalty functions. *Wiley Interdiscip Rev Comput Stat*. 2015;7(1):99–108. [PubMed: 25691921]
11. Fang K, Fan X, Zhang Q, Ma S. Integrative sparse principal component analysis. *J Multivar Anal*. 2018;166:1–16.
12. Guo J, Levina E, Michailidis G, Zhu J. Joint estimation of multiple graphical models. *Biometrika*. 2011;98(1):1–15. [PubMed: 23049124]
13. Danaher P, Wang P, Witten D. The joint graphical lasso for inverse covariance estimation across multiple classes. *J Royal Stat Soc Ser B Stat Methodol*. 2014;76(2):373–397.
14. Shi X, Liu J, Huang J, Zhou Y, Shia BC, Ma S. Integrative analysis of high-throughput cancer studies with contrasted penalization. *Genet Epidemiol*. 2014;38(2):144–151. [PubMed: 24395534]
15. Tseng G, Ghosh D, Zhou X. Integrating Omics Data. New York, NY: Cambridge University Press; 2015.
16. Hirose K, Yamamoto M. Sparse estimation via nonconcave penalized likelihood in factor analysis model. *Stat Comput*. 2015;25(5):863–875.
17. Rothman AJ, Levina E, Zhu J. Sparse multivariate regression with covariance estimation. *J Comput Graph Stat*. 2010;19(4):947–962. [PubMed: 24963268]
18. Yin J, Li H. A sparse conditional gaussian graphical model for analysis of genetical genomics data. *Ann Appl Stat*. 2011;5(4):2630–2650. [PubMed: 22905077]
19. Bühlmann P, van de Geer S. Statistics for High-Dimensional Data. Berlin, Germany: Springer; 2011.
20. Ma S, Huang J, Song X. Integrative analysis and variable selection with multiple high-dimensional data sets. *Biostatistics*. 2011;12(4):763–775. [PubMed: 21415015]
21. Hao B, Sun WW, Liu Y, Cheng G. Simultaneous clustering and estimation of heterogeneous graphical models. *J Mach Learn Res*. 2018;18:1–58.
22. Shi X, Zhao Q, Huang J, Xie Y, Ma S. Deciphering the associations between gene expression and copy number alteration using a sparse double laplacian shrinkage approach. *Bioinformatics*. 2015;31(24):3977–3983. [PubMed: 26342102]
23. Tan KM, London P, Mohan K, Lee SI, Fazel M, Witten D. Learning graphical models with hubs. *J Mach Learn Res*. 2014;15:3297–3331. [PubMed: 25620891]
24. Ni IBP, Zakaria Z, Muhammad R, et al. Gene expression patterns distinguish breast carcinomas from normal breast tissues: the Malaysian context. *Pathol Res Pract*. 2010;206(4):223–228. [PubMed: 20097481]

25. Liu R, Graham K, Glubrecht DD, Germain DR, Mackey JR, Godbout R. Association of FABP5 expression with poor survival in triple-negative breast cancer: implication for retinoic acid therapy. *Am J Pathol.* 2011;178(3):997–1008. [PubMed: 21356353]
26. Yu K, Ganesan K, Tan LK, et al. A precisely regulated gene expression cassette potently modulates metastasis and survival in multiple solid cancers. *PLOS Genet.* 2008;4(7):e1000129. [PubMed: 18636107]
27. Shi X, Shen S, Liu J, Huang J, Zhou Y, Ma S. Similarity of markers identified from cancer gene expression studies: observations from GEO. *Brief Bioinform.* 2014;15(5):671–684. [PubMed: 23788798]
28. Witten D, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics.* 2009;10(3):515–534. [PubMed: 19377034]

FIGURE 1.

Scale free case (a) with $M = p/2$. Proposed: chocolate curves ($\lambda_3 = 0$, solid line; $\lambda_3/n = 0.01$, dashed line; $\lambda_3/n = 0.05$ dotted line; $\lambda_3 = 0.1/n$, dot-dash line); JMG: dark curves. Columns 1 to 4 correspond to $p_d = 0, 0.1, 0.3$ and 0.5 . Row 1: ROC curves. Row 2: ER versus FPR. ER, estimation error; FPR, false positive rate; ROC, receiver operating characteristic

FIGURE 2.

Graphs constructed using JMG (row 1) and the proposed approach (row 2). Columns 1 to 3 correspond to datasets GSE5364, GSE22820, and GSE15852

TABLE 1

Scale free case (a) with $M = p/2$: summary statistics on the models of Akaike information criterion (AIC)-selected λ_1 and λ_2

P_d	Method	λ_3/n	b				θ			
			Angle	NR	TPR	FPR	ER	TPR	FPR	MCC
0	JMG	-	-	-	-	-	1.41	0.82	0.16	0.25
			-	-	-	-	(0.07)	(0.03)	(0.01)	(0.01)
	Proposed	0	5.47	1.00	1.00	0.30	0.98	0.86	0.11	0.32
			(0.41)	(0.03)	0.00	(0.04)	(0.05)	(0.02)	(0.00)	(0.01)
		0.01	5.36	1.00	1.00	0.30	0.67	0.86	0.08	0.36
			(0.38)	(0.03)	0.00	(0.05)	(0.06)	(0.03)	(0.01)	(0.03)
		0.05	4.99	1.00	1.00	0.30	0.34	0.83	0.05	0.44
			(0.34)	(0.03)	0.00	(0.04)	(0.02)	(0.03)	(0.00)	(0.02)
		0.1	4.64	1.00	1.00	0.31	0.28	0.78	0.02	0.56
			(0.33)	(0.03)	0.00	(0.04)	(0.02)	(0.03)	(0.00)	(0.02)
0.1	JMG	-	-	-	-	-	1.26	0.80	0.14	0.26
			-	-	-	-	(0.05)	(0.02)	(0.00)	(0.01)
	Proposed	0	5.31	0.99	1.00	0.29	0.94	0.85	0.11	0.33
			(0.30)	(0.02)	0.00	(0.06)	(0.04)	(0.02)	(0.00)	(0.01)
		0.01	5.21	0.99	1.00	0.28	0.64	0.84	0.08	0.37
			(0.30)	(0.02)	0.00	(0.06)	(0.03)	(0.02)	(0.00)	(0.01)
		0.05	4.86	0.99	1.00	0.29	0.37	0.82	0.05	0.44
			(0.28)	(0.02)	0.00	(0.06)	(0.02)	(0.03)	(0.00)	(0.02)
		0.1	4.51	0.99	1.00	0.28	0.31	0.75	0.02	0.55
			(0.28)	(0.03)	0.00	(0.05)	(0.01)	(0.03)	(0.00)	(0.02)
0.3	JMG	-	-	-	-	-	1.31	0.79	0.16	0.26
			-	-	-	-	(0.05)	(0.02)	(0.01)	(0.01)
	Proposed	0	5.33	0.98	1.00	0.29	0.92	0.81	0.10	0.33
			(0.26)	(0.03)	0.00	(0.04)	(0.04)	(0.02)	(0.00)	(0.01)
		0.01	5.23	0.98	1.00	0.29	0.74	0.82	0.10	0.35
			(0.27)	(0.03)	0.00	(0.04)	(0.07)	(0.02)	(0.01)	(0.02)
		0.05	4.86	0.98	1.00	0.30	0.40	0.77	0.05	0.44
			(0.30)	(0.03)	0.00	(0.04)	(0.02)	(0.03)	(0.00)	(0.02)
		0.1	4.50	0.98	1.00	0.30	0.35	0.70	0.02	0.54
			(0.34)	(0.03)	0.00	(0.05)	(0.02)	(0.02)	(0.00)	(0.02)
0.5	JMG	-	-	-	-	-	1.21	0.79	0.16	0.27
			-	-	-	-	(0.03)	(0.02)	(0.00)	(0.01)
	Proposed	0	5.26	0.99	1.00	0.29	0.85	0.82	0.11	0.34
			(0.33)	(0.02)	0.00	(0.04)	(0.04)	(0.02)	(0.00)	(0.02)
		0.01	5.16	0.99	1.00	0.29	0.64	0.81	0.09	0.37
			(0.32)	(0.02)	0.00	(0.04)	(0.10)	(0.03)	(0.02)	(0.03)
		0.05	4.84	0.99	1.00	0.29	0.39	0.78	0.05	0.45

P_d	Method	λ_3/n	b				θ			
			Angle	NR	TPR	FPR	ER	TPR	FPR	MCC
			(0.28)	(0.02)	0.00	(0.04)	(0.02)	(0.03)	(0.00)	(0.02)
		0.1	4.48	0.99	1.00	0.29	0.36	0.70	0.02	0.56
			(0.23)	(0.03)	0.00	(0.03)	(0.02)	(0.03)	(0.00)	(0.03)

Abbreviations: ER, estimation error; FPR, false positive rate; MCC, Matthews correlation coefficient; NR, the mean ratio of $\|\hat{\mathbf{b}}^{(k)}\| / \|\mathbf{b}^{(k)}\|_S$; TPR, true positive rate.

TABLE 2

Summary statistics of real data analysis

		b		θ	
		JMG	Proposed	JMG	Proposed
Identification					
GSE5364	-	72	1564	1473	
GSE22820	-	60	1412	1316	
GSE15852	-	70	937	814	
Overlapping					
GSE5364, GSE22820	-	55	972	904	
GSE5364, GSE15852	-	63	706	606	
GSE22820, GSE15852	-	53	655	562	