HHS Public Access
Author manuscript
Proc Int Conf Data Sci Adv Anal. Author manuscript; available in PMC 2019 October 01.

Published in final edited form as:
Proc Int Conf Data Sci Adv Anal. 2018 October ; 2018: 109–119. doi:10.1109/DSAA.2018.00021.

Generalized Bayesian Factor Analysis for Integrative Clustering with Applications to Multi-Omics Data

Eun Jeong Min[#*], **Changgee Chang**[#†], and **Qi Long**[‡]
Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania Philadelpia, USA

[#] These authors contributed equally to this work.

Abstract

Integrative clustering is a clustering approach for multiple datasets, which provide different views of a common group of subjects. It enables analyzing multi-omics data jointly to, for example, identify the subtypes of diseases, cells, and so on, capturing the complex underlying biological processes more precisely. On the other hand, there has been a great deal of interest in incorporating the prior structural knowledge on the features into statistical analyses over the past decade. The knowledge on the gene regulatory network (pathways) can potentially be incorporated into many genomic studies.

In this paper, we propose a novel integrative clustering method which can incorporate the prior graph knowledge. We first develop a generalized Bayesian factor analysis (GBFA) framework, a sparse Bayesian factor analysis which can take into account the graph information. Our GBFA framework employs the spike and slab lasso (SSL) prior to impose sparsity on the factor loadings and the Markov random field (MRF) prior to encourage smoothing over the adjacent factor loadings, which establishes a unified shrinkage adaptive to the loading size and the graph structure. Then, we use the framework to extend iCluster+, a factor analysis based integrative clustering approach. A novel variational EM algorithm is proposed to efficiently estimate the MAP estimator for the factor loadings. Extensive simulation studies and the application to the NCI60 cell line dataset demonstrate that the propose method is superior and delivers more biologically meaningful outcomes.

Keywords

Generalized Bayesian Factor Analysis; Markov Random Field (MRF); Spike and Slab Lasso (SSL); Variational EM Algorithm; Structural Information; Network Information; Integrative Analysis; Integrative Clustering; High Dimensional Data; Omics Data; NCI60

[*] mineunj@pennmedicine.upenn.edu
[†] changgee@pennmedicine.upenn.edu
[‡] qlong@pennmedicine.upenn.edu

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

I. Introduction

Rapid advances in technologies have led to the generation of massive amounts of multi-omics data. This has led to a great deal of interest in the integrative analysis of multiple datasets. As the multiple datasets provide information on the subjects and/or the features from many different perspectives, the integrative analysis can help us better understand complex biological underpinnings of many diseases and health problems.

Integrative clustering is the integrative analysis approach that aims to cluster the subjects from which multiple datasets have been gathered. By capturing the underlying biological variants governing the variations of all datasets, we are able to obtain more accurate subtyping, uncover new disease subtypes, understand disease progression, or develop tailored treatment and personalized medicine. For example, while the classification of cancers is typically based on the tissue/cell-type origin and pathogens [1], the biological mechanism of diseases actually builds upon several biological molecular processes [2]. By aggregating the information of the molecular processes at different levels, the integrative clustering can identify new disease subtypes whose clinical outcomes are distinct from the traditional cancer subtypes [3, 4] and help us understand better the survival and mortality risk differences.

Several approaches for integrative clustering have been proposed and applied to multi-omics data [1]. Many approaches are based on the low-rank matrix factorization such as the nonnegative matrix factorization [5] and the factor analysis [6, 7, 4]. Kim et al. [8] utilizes the hierarchical structure among the multi-omics data. Kiselev et al. [9] applies the consensus clustering approach to the single-cell RNA-seq data. Shen et al. [6] introduces iCluster, which is a Bayesian factor analysis based integrative clustering approach. In the framework of iCluster, the low-dimensional latent factors are assumed to carry sufficient information about the underlying biological variations among the subjects, and the posterior distribution of the latent factors are considered much less noisy and more reliable than the original data. Therefore, they perform the K-means clustering on the posterior mean of the latent factors, which are imputed by the EM algorithm. iCluster+ [4] extends iCluster allowing for dealing with more data types such as binary and Poisson in addition to Gaussian. Mo et al. [10] proposes the full Bayesian approach of iCluster.

On the other hand, many techniques of incorporating biological structural information have been developed over the past decade. It is well-known that gene expressions are governed by the gene regulatory network (GRN), which includes various pathways. A gene pathway can be represented by a directed acyclic graph (DAG) where the nodes represent pathway members (genes) and the edges represent production dependencies between the gene pairs. The knowledge on such graphical structure of genes can potentially improve statistical analyses of gene expression data, and is constantly growing fed by various data sources [11]. Some pathway databases are publicly available [12].

The existing statistical approaches capable of incorporating graph information include Li and Li [13] and Pan et al.[14], which propose network-based penalties from the frequentist perspectives. In the Bayesian framework, Li and Zhang [15] and Stingo et al. [16] use the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

spike and slab prior in combination with the Markov random field (MRF) prior. Chang et al. [17] and Rockova and Lesaffre [18] incorporate graph information by smoothing shrinkage parameters under Bayesian lasso framework. The basic principle behind these works is that the penalties or the shrinkage priors encourage group-wise selection of features that are adjacent on the graph structure or belong to a same pathway. In addition to the biological qualitative motivation offered by the GRN, Yu and Liu [19] provides a quantitative justification of the group-wise selection. However, all of the aforementioned approaches are developed for the regression framework. For clustering, we note that the PARADIGM [20] uses the pathway level activities instead of the individual gene expression data, but no clustering method has been proposed to utilize the pathway graph information, to the best of our knowledge.

In this paper, we propose a novel integrative clustering method that can incorporate the graph information. We first build the generalized Bayesian factor analysis (GBFA) framework, and use the framework to extend iCluster+ [4]. Our GBFA framework employs the spike and slab lasso (SSL) prior to impose sparsity on factor loadings and the MRF prior [15] or the Ising prior to incorporate network information. The combination of the SSL and MRF (or Ising) priors achieves doubly adaptive L_1 shrinkage on factor loadings; the level of shrinkage is adaptive to the corresponding loading size and its neighboring shrinkage levels. The adaptivity to loading sizes alleviates the bias caused by the L_1 shrinkage and the adaptivity to neighboring shrinkage parameters encourages the group-wise selection among the adjacent factor loadings, which are the first innovation of our work. Plus, the SSL prior allows the Maximum a Posteriori (MAP) estimates of the factor loadings to attain exact zeros unlike the traditional spike and slab Gaussian prior. This leads to more accurate estimation of other parameters involving the latent factors, which the subsequent clustering procedure will be based on.

Similar approaches have been proposed by Klami et al. [21], Virtanen et al. [22], which they call the group factor analysis. Note that, while their approaches incorporate the group membership information, our GBFA takes into account the network graph information, which contains finer interaction knowledge beyond the membership. Also, while their approaches can deal with continuous variables only, our GBFA can support multiple data types including discrete variables such as binomial, negative binomial, and Poisson. This is the second innovation of our work.

The third innovation is that we propose an efficient variational EM algorithm for the MAP estimator of the factor loadings, where the posterior mean and covariance of the latent factors are imputed by minimizing the Kullback-Leibler divergence to the original posterior distribution. Owing to the latent variable augmentation [23], all relevant expectations are completely tractable, and therefore we can avoid the Monte Carlo approximation used in [4]. The imputed posterior mean of the latent factors are then used for clustering via the K-means clustering method.

The paper is organized as follows. In the following section, we propose the GBFA framework and the integrative clustering method. Section III presents the variational EM algorithm. Extensive simulation studies are conducted in Section IV, followed by the

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

application to the NCI60 dataset in Section V. Section VI includes conclusions and discussion.

II. Methodology

In this section, we first establishes the generalized Bayesian factor analysis framework. Then, we propose an integrative clustering algorithm based on the GBFA framework.

A. Notation

Vectors are denoted by bold and lowercase letters. Matrices are denoted by uppercase letters if not greek. For a $p \times n$ matrix A, a_{ji} stands for the (j, i) entry of A, $\mathbf{a}_i = (a_{1i},..., a_{pi})^T$ stands for the i-th column vector of A, and $\tilde{\mathbf{a}}_j = (a_{j1},..., a_{jn})^T$ stands for the j-th row vector of A. For a vector b, b_i stands for the i-th element of b and D_{b} denotes the diagonal matrix diag(b).

B. Generalized Bayesian Factor Analysis Framework

Suppose we have H multi-modal data generated from various technologies, say X^1, ..., X^H, each of which is a $ph \times n$ matrix for $1 \le h \le H$ where n is the sample size. Let X be their vertical concatenate, which is of size $p \times n$ where $p = \sum_{h=1}^{H} ph$.

$$X = \begin{bmatrix} X^1 \\ \vdots \\ X^H \end{bmatrix} \in \mathbb{R}^{p \times n}.$$

Let $\mathbf{z}_i$ be the L-dimensional latent factor for the i-th subject and $\tilde{\mathbf{w}}_j$ be the factor loadings for the j-th feature. Let $\mu = \mathrm{m}1^{\mathrm{T}} + WZ$ where m is the location vector, $Z = [\mathbf{z}_1,...,\mathbf{z}_n] \in \mathbb{R}^{L\times n}$ be the latent factor matrix, and $W = [\tilde{\mathbf{w}}_1,..., \tilde{\mathbf{w}}_{\mathrm{p}}]^{\mathrm{T}} \in \mathbb{R}^{p\times L}$ be the factor loading matrix. GBFA assumes that the distribution of X is governed by the parameter matrix μ, which takes a low-rank representation, i.e., $L \ll p$.

Suppose $\{x_{ji}\}_{1\le i\le n}$ belong to the same distribution family and the distribution of X, parameterized by μ, has the following form of likelihood.

$$\pi(X \mid \mu) = \prod_j \prod_i \pi_j(x_{ji} \mid \mu_{ji}),$$

where π_j is the density function for the j-th distribution family.

Note that it is critical to support various and heterogeneous data types in accommodating multi-omics data. Our framework allows for analyzing heterogeneous data types including binomial, negative binomial, Poisson, and Gaussian distributions, and is more flexible than iCluster+ [4], which admits Bernoulli, Poisson, and Gaussian distributions only. Suppose $\{x_{ji}\}_{1\le i\le n}$ follow the binomial distribution with parameters n_j and p_{ji}. We use the logit link function for p_{ji} and the likelihood is given by

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

$$\pi_j(x_{ji} \mid \mu_{ji}) = \binom{n_j}{x_{ji}} \frac{e^{\mu_{ji} x_{ji}}}{(1 + e^{\mu_{ji}})^{n_j}}, \; 0 \le x_{ji} \le n_j . \quad (1)$$

If $\{x_{ji}\}_{1 \le i \le n}$ follow the negative binomial distribution with parameters r_j and p_{ji}, for which we use the logit link function, the likelihood is given by

$$\pi_j(x_{ji} \mid \mu_{ji}) = \binom{r_j + x_{ji} - 1}{x_{ji}} \frac{e^{\mu_{ji} x_{ji}}}{(1 + e^{\mu_{ji}})^{r_j + x_{ji}}}, \; x_{ji} \ge 0 . \quad (2)$$

If $\{x_{ji}\}_{1 \le i \le n}$ follow the Poisson distribution with mean $e^{\mu_{ji}}$, the likelihood is given as follows.

$$\pi_j(x_{ji} \mid \mu_{ji}) = e^{-e^{\mu_{ji}}} \frac{e^{\mu_{ji} x_{ji}}}{x_{ji}!}, \; x_{ji} \ge 0 . \quad (3)$$

If $\{x_{ji}\}_{1 \le i \le n}$ follow the Gaussian distribution with mean μ_{ji} and precision ρ_j, we have the likelihood as follows.

$$\pi_j(x_{ji} \mid \mu_{ji}) = \frac{\rho_j^{1/2}}{\sqrt{2\pi}} e^{-\rho_j (x_{ji} - \mu_{ji})^2 / 2}, x_{ji} \in \mathbb{R} .$$

Binomial, negative binomial, and Gaussian distributions have extra parameters n_j, r_j, and ρ_j, respectively. The parameters n_j and r_j must be provided with the data. We assign the gamma distribution $\mathcal{G}\left(\zeta_j/2, \zeta_j/2\right)$ as the prior for ρ_j. A typical choice for ς_j is 1, which gives the most uninformative prior. We marginalize out ρ_j and the actual (marginal) distribution of x_{ji} becomes the Student *t*-distribution.

Of note, the Bernoulli distribution is a special case of the binomial distribution with $n_j = 1$ and the geometric distribution is a special case of the negative binomial distribution with $r_j = 1$.

C. Priors for W and Z

We consider the L_1 shrinkage spike and slab prior on W,

$$\log \pi(W \mid \gamma) = C + \sum_{j,l} \log \lambda_{jl} - \sum_{j,l} \lambda_{jl} \mid w_{jl} \mid ,$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

where $\lambda_{jl} = (1 - \gamma_{jl})\lambda_0 + \gamma_{jl}\lambda_1$ with $0 \leq \lambda_1 < \lambda_0$ and γ_{jl} is a binary variable indicating nonzero-ness of w_{jl}. This spike and slab lasso (SSL) prior, which has been introduced in [24, 25], administers two levels of shrinkage depending on the size of loadings. If $|w_{jl}|$ is close to zero, γ_{jl} turns off and triggers higher level of shrinkage λ_0 for w_{jl}. Conversely, if $|w_{jl}|$ is far from zero, γ_{jl} turns on and w_{jl} receives lower level of shrinkage λ_1. Furthermore, combined with EM algorithm, the SSL prior imposes continuously adaptive shrinkage on w_{jl}, as elaborated in Section III.

An important advantage of SSL prior against the traditional spike and slab Gaussian prior [26] is that the elements in the MAP estimator $\hat{W}$ can take exact zero values. Having exact zero values for true zeros in W leads to more accurate estimation of other parameters including the ones related to the latent factors Z. As the clustering algorithm, presented in Section II-E, is based on the posterior mean of Z, it makes a crucial improvement.

Following the standard orthonormality assumption on the latent factors, we assign the standard Gaussian prior for Z.

$$\log \pi(Z) = C - \frac{1}{2}\sum_{l,i} z_{li}^2 .$$

D. Prior for γ: Incorporating Network Information

Suppose the graphs $G_h = \langle P_h, E_h \rangle$ are given where $P_h = \{1,...,p_h\}$ represents the set of variables in the h-th dataset and E_h is the set of edges among them, where the presence of edge indicates the correlation between the relevant pair of variables. We incorporate this network information into the generalized factor analysis framework by encouraging to match the correlation structure. Since X^h are vertically concatenated, we first combine H graphs into a single graph $G = \langle P, E \rangle$ by setting $P = \{1, ..., p\}$ and $E = \{(\imath(h,j), \imath(h,k)) : (j, k) \in E_h, 1 \leq h \leq H\}$ where $\imath(h,j)$ is the index in the matrix X of the j-th variable in the h-th dataset. Let G be the adjacency matrix for G.

Consider two Gaussian variables x_j and x_k. Note that, under the assumption that the latent factors z_{li} are independent with zero mean and unit variance, the covariance of x_j and x_k given W is as follows.

$$\mathrm{Cov}(x_j, x_k) = \sum_l w_{jl} w_{kl} .$$

This implies that, in order for x_j and x_k to be correlated, they must load at least one common factor. This argument easily extends to non-Gaussian variables as well. If the latent factors are independent, the only way a pair of variables can be correlated is that they must load at least one common factor.

Therefore, it is reasonable to encourage the pairs of adjacent variables to share common factors. If one variable loads any factor, its adjacent variables are encouraged to load the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

factor as well. In other words, if x_j and x_k are adjacent in G and if $w_{jl} \neq 0$ for some l, then we promote $w_{kl} \neq 0$.

To this end, we employ the Markov random field (MRF) prior, as introduced in Li and Zhang [15].

$$\log \pi(\gamma) = C_{\delta,\eta} - \delta \sum_{j,l} \gamma_{jl} + \eta \sum_{j,k,l} G_{jk} \gamma_{jl} \gamma_{kl}, \quad (4)$$

where δ is the parameter that controls of the sparsity of factor loadings. The MRF prior helps achieve our goal because $\gamma_{kl} = 1$ is encouraged if $\gamma_{jl} = 1$ and vice versa, when x_j and x_k are adjacent ($G_{jk} = 1$). Here, $\eta \geq 0$ is the smoothing parameter that controls the strength of the encouragement.

We also propose using the Ising prior for γ_{jl} as follows.

$$\log \pi(\gamma) = C_{\delta,\eta} - \delta \sum_{j,l} \gamma_{jl} + \eta \sum_{j,k,l} G_{jk} \mathbb{I}(\gamma_{jl} = \gamma_{kl}) . \quad (5)$$

The key difference between (4) and (5) is that the former encourages $\gamma_{jl} = \gamma_{kl} = 1$ only while the latter promotes $\gamma_{jl} = \gamma_{kl}$. The two-way smoothing in (5) can enhance specificity depending on the specified graph structure.

E. Clustering

In the GBFA framework, the high-dimensional data matrix X is represented by the low-rank parameter matrix μ. In particular, the low-dimensional latent factors Z are shared by all data sets and assumed to capture all biological variations among the subjects. Since Z is not observed, we retrieves the cluster membership based on the posterior mean $\mathbb{E}(Z|X)$. When the model assumptions are true, $\mathbb{E}(Z|X)$ is considered much less noisy and more reliable than the original data matrix X. In Section III, we develop the variational EM algorithm, where $\mathbb{E}(Z|X)$ is approximated by μ_Z. We then conduct the K-means clustering on the estimate of μ_Z to retrieve the membership of clusters. We call the resulting clustering method GBFA_CL.

III. Computation

We propose the Maximum A Posteriori (MAP) estimator for W marginalizing Z and γ out. Unfortunately, the classical EM approach involves an intractable conditional expectation of the loglikelihood. To address the problem, we use the latent variable augmentation technique and the variational EM approach. For binomial and negative binomial likelihood functions, we use the following identity [23].

$$\frac{e^{\mu_{ji} x_{ji}}}{(1 + e^{\mu_{ji}})^{b_{ji}}} = 2^{-b_{ji}} e^{\kappa_{ji} \mu_{ji}} \int_0^\infty e^{-\rho_{ji} \mu_{ji}^2 / 2} \pi_{ji}(\rho_{ji}) d\rho_{ji},$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

where $k_{ji} = x_{ji} - b_{ji}/2$ and $\boldsymbol{\pi}_{ji}(\rho_{ji})$ is the density of the Polya-Gamma class $\mathscr{P}\mathscr{G}\left(b_{ji}, 0\right)$. Applying this identity to (1), (2), and (3) offers a unified form of likelihood for Gaussian, binomial, and negative binomial variables as follows.

$$\pi_j(\widetilde{\mathbf{x}}_j, \widetilde{\boldsymbol{\rho}}_j \mid \widetilde{\boldsymbol{\mu}}_j) \propto e^{-\frac{1}{2}\sum_i \rho_{ji}(\mu_{ji} - \psi_{ji})^2 + \sum_i \kappa_{ji}\mu_{ji}} \pi_j^*(\widetilde{\boldsymbol{\rho}}_j), \quad (6)$$

where the unknown components are summarized in Table I. There is no ρ_{ji} associated with a Poisson variable x_{ji}. But, for succinct notations, we use $\pi_j(\widetilde{\mathbf{x}}_j, \widetilde{\boldsymbol{\rho}}_j \mid \widetilde{\boldsymbol{\mu}}_j) = \pi_j(\widetilde{\mathbf{x}}_j \mid \widetilde{\boldsymbol{\mu}}_j)\pi_j^*(\widetilde{\boldsymbol{\rho}}_j)$ with any π_j^*. Then, the full likelihood can be written as

$$\pi(\boldsymbol{\gamma}, \boldsymbol{\rho}, W, Z, X \mid \mathbf{m}) = \pi(W \mid \boldsymbol{\gamma})\pi(\boldsymbol{\gamma})\pi(Z)\prod_j \pi_j(\widetilde{\mathbf{x}}_j, \widetilde{\boldsymbol{\rho}}_j \mid \widetilde{\boldsymbol{\mu}}_j).$$

A. Variational EM Algorithm

Let $\tilde{\boldsymbol{\pi}}(\cdot)$ be a measure on $\boldsymbol{\gamma}$, $\boldsymbol{\rho}$, and Z and $\widetilde{\mathbb{E}}(\cdot)$ denotes the expectation with respect to $\tilde{\boldsymbol{\pi}}$. The EM algorithm yields the MAP estimator by maximizing

$$Q\left(\mathbf{m}, W, \widetilde{\pi}\right) = \widetilde{\mathbb{E}} \log \pi\left(\boldsymbol{\gamma}, \boldsymbol{\rho}, W, Z, X \mid \mathbf{m}\right) - \widetilde{\mathbb{E}} \log \widetilde{\pi}\left(\boldsymbol{\gamma}, \boldsymbol{\rho}, Z\right),$$

which is maximized when $\tilde{\boldsymbol{\pi}}(\cdot) = \boldsymbol{\pi}(\cdot|\mathrm{m}, W, X)$ and $Q(\mathrm{m}, W, \tilde{\boldsymbol{\pi}}) = \log \boldsymbol{\pi}(W, X|\mathrm{m})$.

Unfortunately, there is no analytic solution to the conditional expectations involving $\boldsymbol{\gamma}$, $\boldsymbol{\rho}$, and Z given W. Therefore, we use the variational EM approach [27], which limits the domain of $\tilde{\boldsymbol{\pi}}$ within tractable. We consider a product measure on individual γ_{jl}'s, $\boldsymbol{\rho}$, and Z. Let $\hat{\boldsymbol{\pi}}(\boldsymbol{\gamma}, \boldsymbol{\rho}, Z) = \hat{\boldsymbol{\pi}}(\boldsymbol{\gamma})\hat{\boldsymbol{\pi}}(\boldsymbol{\rho})\hat{\boldsymbol{\pi}}(Z)$ where

$$\begin{aligned}
\widehat{\pi}(\boldsymbol{\gamma}) &= \prod_{j,l} \theta_{jl}^{\gamma_{jl}}(1 - \theta_{jl})^{1 - \gamma_{jl}}, \\
\widehat{\pi}(\boldsymbol{\rho}) &\propto \prod_j e^{-\frac{1}{2}\sum_i \rho_{ji}\varphi_{ji}^2} \pi_j^*(\widetilde{\boldsymbol{\rho}}_j), \\
\widehat{\pi}(Z) &\propto \prod_i e^{-\frac{1}{2}(\mathbf{z}_i - \boldsymbol{\mu}_{z,i})^T \Sigma_{z,i}^{-1}(\mathbf{z}_i - \boldsymbol{\mu}_{z,i})},
\end{aligned}$$

and $\hat{\mathbb{E}}$ be the expectation operator under $\hat{\boldsymbol{\pi}}$. Note that $\{\rho_{ji}\}_{1 \le i \le n}$ associated with a Poisson variables x_j are completely independent, for which we freely fix $\varphi_{ji} = 0$.

Then, our MAP estimator is obtained by maximizing the evidence lower bound (ELBO) with respect to m, W, $\boldsymbol{\theta}$, $\boldsymbol{\varphi}$, and $\{\mu_{z,i}, \sum_{z,\mathrm{i}}\}_{1 \le i \le n}$

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

$$\begin{aligned} Q(\mathbf{m}, W, \boldsymbol{\theta},\ \boldsymbol{\varphi}, \{\boldsymbol{\mu}_{z,i}, \Sigma_{z,i}\}_{i=1}^{n}) &= \widehat{\mathbb{E}} \log \pi(\boldsymbol{\gamma}, \boldsymbol{\rho}, W, Z, X \mid \mathbf{m}) \\ &-\widehat{\mathbb{E}} \log \widehat{\pi}(\boldsymbol{\gamma}) - \widehat{\mathbb{E}} \log \widehat{\pi}(\boldsymbol{\rho}) - \widehat{\mathbb{E}} \log \widehat{\pi}(Z), \end{aligned} \tag{7}$$

which lower-bounds the log evidence log $\boldsymbol{\pi}(W, X|\mathrm{m})$.

Each of E- and M-steps will be detailed in the following sections. A single iteration of our EM algorithm takes $O(npL^2 + eL)$ time complexity where $e = |E|$ is the number of edges in the graph, assuming $L \ll \min(p, n)$. The EM approach has an advantage over MCMC in that the number of iterations that EM algorithm requires is generally much smaller than the number of samples that MCMC needs. It is worth noting that E-step for Z and M-step for W are the bottleneck but can be obviously parallelized. Finally, we note that any coordinate ascent algorithm would be inappropriate because the objective function is not concave and the solution will depend on the order of the variables.

B. E-steps

E-steps optimize (7) with respect to the parameters of $\hat{\boldsymbol{\pi}}(Z)$, $\hat{\boldsymbol{\pi}}(\boldsymbol{\rho})$, and $\hat{\boldsymbol{\pi}}(\boldsymbol{\gamma})$, alternately.

1) E-step for $\boldsymbol{\gamma}$:

Let $\alpha_{jl} = \log\frac{\theta_{jl}}{1-\theta_{jl}}$. We make a quasi-Newton update for $\boldsymbol{\alpha}_l$ with backtracking line search for each l. As the Hessian matrix is huge if p is large, we make use of only the diagonal elements of the Hessian matrix following Becker and Le Cun [28]. This alleviates the gradient vanishing problem of the sigmoid (expit) function.

As alluded to in Section I, we can see that the SSL and MRF priors indeed achieve the doubly adaptive shrinkage for W. The KKT condition yields, for each α_{jl},

$$\alpha_{jl} = -\delta + \eta \sum_{k} G_{jk}\theta_{kl} + \log\frac{\lambda_1}{\lambda_0} - |w_{jl}|(\lambda_1 - \lambda_0). \tag{8}$$

Section III-C1 and (8) show that L_1 shrinkage for w_{jl} depends on the loading size $|w_{jl}|$ and the shrinkage for adjacent loadings (θ_{kl}).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Algorithm 1: EM algorithm for GBFA

```
$P \leftarrow \{j: x_j$ is Poisson$\}$.
Initialize $\mathbf{m}$, $W$, $\rho$, and $\theta$.

repeat
    for $l = 1, \ldots, L$ do                                          $/ \star$ E − step for $\boldsymbol{\gamma} \star /$
        $\alpha_{jl} \leftarrow \alpha_{jl} + s_l(-\delta + \eta \sum_k G_{jk}\theta_{kl} + \log(\lambda_1 / \lambda_0) - |w_{jl}|(\lambda_1 - \lambda_0) - \alpha_{jl})$ for all $j$ where $s_l$ is determined by the backtracking line search.
        $\theta_{jl} \leftarrow 1 / (1 + e^{-\alpha_{jl}})$; $\lambda_{jl} \leftarrow (1 - \theta_{jl})\lambda_0 + \theta_{jl}\lambda_1$ for all $j$.
    end
    if $\mathscr{P} = \varnothing$ then update $\boldsymbol{\mu}_{z,i}$ and $\Sigma_{z,i}$ as in (9) for $1 \le i \le n$.     $/ \star$ E − step for $Z \star /$
    else
        for $i = 1, \ldots, n$ do update $\boldsymbol{\mu}_{z,i}$ and $\Sigma_{z,i}$ by the gradient ascent method with the backtracking line search.
    end
    for $j = 1, \ldots, p$ do update $\varphi_{ji}$, $\rho_{ji}$, and $\rho_j$ as in Section III-B2.     $/ \star$ E − step for $\boldsymbol{\rho} \star /$
    for $j = 1, \ldots, p$ do                                          $/ \star$ M − step for $W \star /$
        if $j \notin \mathscr{P}$ then $\widetilde{\mathbf{w}}_j \leftarrow \operatorname{argmin}_{\mathbf{w}}(\frac{1}{2}\mathbf{w}^T A_j \mathbf{w} - \mathbf{b}_j^T \mathbf{w} + \sum_l \lambda_{jl}|w_l|$ where $A_j$ and $\mathbf{b}_j$ are given by (10).
        else update $\widetilde{\mathbf{w}}_j$ by the proximal gradient ascent method.
    end
until convergence.
```

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

2) E-step for $\boldsymbol{\rho}$: For a non-Poisson variable x_j, we have

$$\varphi_{ji}^2 \leftarrow (m_j + \widetilde{\mathbf{w}}_j^T \boldsymbol{\mu}_{z,i} - \psi_{ji})^2 + \widetilde{\mathbf{w}}_j^T \Sigma_{z,i} \widetilde{\mathbf{w}}_j .$$

For a non-Poisson discrete variable x_j, we have by Poison et al. [23]

$$\widehat{\mathbb{E}}(\rho_{ji}) = \frac{b_{ji}(e^{\varphi_{ji}} - 1)}{2\varphi_{ji}(e^{\varphi_{ji}} + 1)}, \quad 1 \le i \le n,$$

with $b_{ji} = n_j$ if x_j is a binomial variable and $b_{ji} = x_{ji} + r_j$ if x_j is a negative binomial variable. For a Gaussian variable x_j, we have

$$\widehat{\mathbb{E}}(\rho_j) = \frac{\zeta_j + n}{\zeta_j + \sum_i \varphi_{ji}^2} .$$

3) E-step for Z: Exact solutions exist for $\mu_{z,i}$ and $\sum_{z,i}$ if there is no Poisson variable in data. For each i, we update

$$\Sigma_{z,i} \leftarrow (W^T \mathscr{D}_{\boldsymbol{\rho}_i} W + I)^{-1}, \boldsymbol{\mu}_{z,i} \leftarrow \Sigma_{z,i} W^T \mathbf{c}_i, \quad (9)$$

where $C = \boldsymbol{\kappa} + \boldsymbol{\rho} \bigcirc (\psi - \mathbf{m1}^T)$. Here, $\bigcirc$ denotes the Hadamard product.

The gradient ascent updates with backtracking line search are performed if at least one Poisson variable is included in data.

C. M-steps

M-steps optimize (7) with respect to W (and $\mathbf{m}$ jointly).

M-step for W: For a non-Poisson variable x_j, we solve

$$\widetilde{\mathbf{w}}_j \leftarrow \operatorname*{argmin}_{\mathbf{w}} \left(\frac{1}{2} \mathbf{w}^T A_j \mathbf{w} - \mathbf{b}_j^T \mathbf{w} + \sum_l \widehat{\mathbb{E}}(\lambda_{jl}) \mid w_l \right),$$

where $\widehat{\mathbb{E}}(\lambda_{jl}) = \theta_{jl} \lambda_1 + (1 - \theta_{jl}) \lambda_0$, and

Author Manuscript

$$
\begin{aligned}
A_j &= -\frac{d^2 \widehat{\mathbb{E}} \log \pi_j(\widetilde{\mathbf{x}}_j, \widetilde{\boldsymbol{\rho}}_j \mid \widetilde{\boldsymbol{\mu}}_j)}{d\widetilde{\mathbf{w}}_j d\widetilde{\mathbf{w}}_j^T}, \\
\mathbf{b}_j &= A_j \widetilde{\mathbf{w}}_j + \frac{d \widehat{\mathbb{E}} \log \pi_j(\widetilde{\mathbf{x}}_j, \widetilde{\boldsymbol{\rho}}_j \mid \widetilde{\boldsymbol{\mu}}_j)}{d\widetilde{\mathbf{w}}_j}.
\end{aligned}
\qquad (10)
$$

Note that A_j and b_j do not depend on the current value of $\tilde{\mathrm{w}}_j$, and this is an adaptive lasso problem where the penalties are adaptive to both the corresponding loading sizes and the graph structure, as so are θ_{jl}s. We use the DWL algorithm [29], which efficiently computes the exact solution. This guarantees the monotone increase of Q.

For a Poisson variable x_j, $\tilde{\mathrm{w}}_\mathrm{j}$ cannot be solved by lasso. Instead, we use the proximal gradient ascent method.

2) M-step for m (jointly with W): If m_j is fitted rather than fixed, we update m_j and $\tilde{\mathrm{w}}_\mathrm{j}$ jointly. For a non-Poisson variable x_j, we can update $(m_j, \tilde{\mathrm{w}}_\mathrm{j})^T$ with extended A_j and $\mathbf{b}_j$ as follows.

$$
A_j = -\begin{bmatrix} \dfrac{d^2 \widehat{\mathbb{E}} \log \pi_j(\widetilde{\mathbf{x}}_j, \widetilde{\boldsymbol{\rho}}_j \mid \widetilde{\boldsymbol{\mu}}_j)}{dm_j^2} & \dfrac{d^2 \widehat{\mathbb{E}} \log \pi_j(\widetilde{\mathbf{x}}_j, \widetilde{\boldsymbol{\rho}}_j \mid \widetilde{\boldsymbol{\mu}}_j)}{dm_j d\widetilde{\mathbf{w}}_j^T} \\ \dfrac{d^2 \widehat{\mathbb{E}} \log \pi_j(\widetilde{\mathbf{x}}_j, \widetilde{\boldsymbol{\rho}}_j \mid \widetilde{\boldsymbol{\mu}}_j)}{d\widetilde{\mathbf{w}}_j dm_j} & \dfrac{d^2 \widehat{\mathbb{E}} \log \pi_j(\widetilde{\mathbf{x}}_j, \widetilde{\boldsymbol{\rho}}_j \mid \widetilde{\boldsymbol{\mu}}_j)}{d\widetilde{\mathbf{w}}_j d\widetilde{\mathbf{w}}_j^T} \end{bmatrix},
$$

$$
\mathbf{b}_j = A_j \begin{bmatrix} m_j \\ \widetilde{\mathbf{w}}_j \end{bmatrix} + \begin{bmatrix} \dfrac{d \widehat{\mathbb{E}} \log \pi_j(\widetilde{\mathbf{x}}_j, \widetilde{\boldsymbol{\rho}}_j \mid \widetilde{\boldsymbol{\mu}}_j)}{dm_j} \\ \dfrac{d \widehat{\mathbb{E}} \log \pi_j(\widetilde{\mathbf{x}}_j, \widetilde{\boldsymbol{\rho}}_j \mid \widetilde{\boldsymbol{\mu}}_j)}{d\widetilde{\mathbf{w}}_j} \end{bmatrix}.
$$

Note that m_j receives no shrinkage. Again, for a Poisson variable x_j, $\tilde{\mathrm{w}}_j$ cannot be solved by lasso. Instead, we use the proximal gradient ascent method.

D. Initialization

Let Y be a matrix approximating μ, which is defined as

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$y_{ji} = \begin{cases} x_{ji}, & \text{if } x_j \text{ is Gaussian}, \\ \operatorname{logit}\left(\dfrac{x_{ji}+1}{n_j+2}\right), & \text{if } x_j \text{ is binomial}, \\ \operatorname{logit}\left(\dfrac{x_{ji}+1}{r_j+x_{ji}+2}\right), & \text{if } x_j \text{ is negative binomial}, \\ \log(x_{ji} \vee 1/2), & \text{if } x_j \text{ is Poisson}. \end{cases}$$

For the location m_j, one can choose the (trimmed) mean of $\tilde{y}_j$, the median of $\tilde{y}_j$, or any fixed vector. One way to initialize W would be the (reduced) singular value decomposition of $Y - \mathbf{m}\mathbf{1}^T = UDV^T$ and $W = UD$. We initialize $\varphi_{ji} = 0$ and $\theta_{jl} = 0.5$. The algorithm is summarized in Algorithm 1.

IV. Simulation Study

We conduct simulation study to evaluate performance of our proposed method. We compare its performance to other methods; the K-means clustering applied to separate and concatenated data, and iCluster+ [4].

A. Simulation Design

We generate 100 Monte Carlo datasets. For each dataset, we generate $H = 3$ data matrices mimicking multi-omics data, each of which has $p_1 = p_2 = p_3 = 300$ features of homogeneous data types varying with h. For each h, we have $\mu^h = \mathbf{m}^h\mathbf{1}^T + W^hZ$ where $\mu^h \in \mathbb{R}^{Ph\times n}$, $\mathbf{m}^h \in \mathbb{R}^{Ph\times 1}$, $W^h \in \mathbb{R}^{Ph\times L}$, and $Z \in \mathbb{R}^{L\times n}$.

We have $L = 3$ latent factors. The true loading matrix W^h is generated as follows.

$$w_{jl}^1 = \begin{cases} \mathscr{U}(1.5, 2.5), & \text{if } \{l = 1, j = 1, \dots, 20\}, \\ & \text{or } \{l = 2, j = 281, \dots, 300\}, \\ 0, & \text{otherwise}, \end{cases}$$
$$w_{jl}^2 = \begin{cases} \mathscr{U}(1.5, 2.5), & \text{if } \{l = 1, j = 281, \dots, 300\}, \\ & \text{or } \{l = 3, j = 1, \dots, 20\}, \\ 0, & \text{otherwise}, \end{cases}$$
$$w_{jl}^3 = \begin{cases} \mathscr{U}(1.5, 2.5), & \text{if } \{l = 2, j = 1, \dots, 20\}, \\ & \text{or } \{l = 3, j = 281, \dots, 300\}, \\ 0, & \text{otherwise}, \end{cases}$$

where U stands for the uniform distribution. All nonzero entries of W^h flip their signs with probability 1/2. Figure 1 illustrates the form of W.

We consider two cases for the true number of clusters; $K = 4$ and $K = 6$. When $K = 4$, there are $n = 160$ subjects. When $K = 6$, there are $n = 180$ subjects. All subjects are grouped to into K equally sized disjoint clusters. If the i-th subject belongs to k-th cluster, the latent factor for the i-th subject is generated as follows.

$$\mathbf{z}_i = \boldsymbol{\nu}_k + \sigma_z \times \mathscr{N}(\mathbf{0}, \mathbf{I}),$$

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

where, when $K = 4$, we use $\boldsymbol{\sigma}_z = 0.3$ and

$$\boldsymbol{\nu}_1 = (1/2, 1/\sqrt{2}, 0)^T, \quad \boldsymbol{\nu}_2 = (1/2, -1/\sqrt{2}, 0)^T,$$
$$\boldsymbol{\nu}_3 = (-1/2, 0, 1/\sqrt{2})^T, \ \boldsymbol{\nu}_4 = (-1/2, 0, -1/\sqrt{2})^T,$$

and when $K = 6$, we use $\boldsymbol{\sigma}_z = 0.2$ and

$$\boldsymbol{\nu}_1 = (1/\sqrt{2}, 0, 0)^T, \ \boldsymbol{\nu}_2 = (-1/\sqrt{2}, 0, 0)^T,$$
$$\boldsymbol{\nu}_3 = (0, 1/\sqrt{2}, 0)^T, \ \boldsymbol{\nu}_4 = (0, -1/\sqrt{2}, 0)^T,$$
$$\boldsymbol{\nu}_5 = (0, 0, 1/\sqrt{2})^T, \ \boldsymbol{\nu}_6 = (0, 0, -1/\sqrt{2})^T.$$

We consider three scenarios, each of which has a different combination of data types as described in Table II. Any Gaussian variable is sampled from $\mathcal{N}(\mu_{ji}^h, \sigma_e^2)$. When $K = 4$, we use $\boldsymbol{\sigma}_\mathrm{e} = 3$, and when $K = 6$, we use $\boldsymbol{\sigma}_\mathrm{e} = 2$. Any binary variable has the success probability $p_{ji}^h = 1 \Big/ (1 + \exp(-\mu_{ji}^h))$. Any Poisson variable has the mean $\exp(-\mu_{ji}^h)$ with $m_j^h = 2.5$. For any binomial variable, the number of trial parameter n_j is randomly selected between 1 and 15 and the success probability is given by $p_{ji}^h = 1 \Big/ (1 + \exp(-\mu_{ji}^h))$. We use $m_j^h = 0$ for all data types except for Poisson.

In all three scenarios, we assume that every 20 consecutive genes in X^1 and X^2 form a pathway and we randomly generate 50 edges within each pathway.

B. Tuning Strategy and Performance Measure

To select the GBFA model, we use the Bayesian information criteria (BIC). We select the model that gives the minimum value of the BiC defined as follows.

$$\mathrm{BIC} = -2\sum_h l_h(X^h, \hat{\mu}^h) + \log(n) \times df,$$

where l_h is the log-likelihood of h-th data and df is the number of nonzero elements in $\hat{W}$.

For a given η, we select the tuning parameter combination (L, λ_0, δ) that gives the minimum BIC value. In all simulations, we use the Ising prior for smoothing, and fix $\lambda_1 = \lambda_0/5$ and **m** at the trimmed mean of $\tilde{y}_j$ (see Section III-D).

We conduct the K-means clustering on the estimated posterior mean μ_Z of the latent factors. To select the optimal number of clusters, there are several widely used methods such as the elbow method [30], the silhouette method [31], gap statistics [32], and so on. We use the silhouette analysis for our proposed method and the K-means clustering method.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

For iCluster+, we use the R package iClusterPlus and perform the most extensive search to choose the best tuning parameters and the best number of clusters K (and of latent factors $L = K - 1$), as suggested by [4].

In order to measure the similarity between the estimated clustering and the true clustering, we use the Jaccard index [33] and the Rand index [34], both of which have values between 0 and 1. The higher the value is, the more similar the clusterings are.

C. Results

The summary of all simulation results is shown in Table III. In all aspects of performance measures, GBFA_CL outperforms other methods in all scenarios. In particular, the ability to choose the right number of clusters of the proposed method is by far superior to that of other methods.

GBFA_CL we obtains improved results when $\eta \neq 0$ in general compared to when $\eta = 0$, which proves the usefulness of the ability to incorporate the network graph information in the GBFA framework. However, the observed improvement is subtle when the result is already very good without incorporation of the graph information.

Note that the concatenated K-means clustering is not necessarily better than the separate K-means clusterings. In fact, we observe each separate K-means clustering works better in many cases. This is due to the fact that the errors are accumulated in the concatenated data, and suggests that a naïve concatenation of data is not a good practice of the integrative analysis. On the other hand, GBFA_CL and iCluster+ outperform the K-means clustering, which demonstrates the effectiveness of the low-dimensional factor analysis when the model assumptions are applicable.

It is surprising that GBFA_CL substantially outperforms iCluster+ even when the graph information is not incorporated. Our understanding is that it is largely owing to our variational EM algorithm being accurate and efficient. Of note, iCluster+ uses Monte Carlo Newton-Raphson type of algorithm due to intractability of the relevant expectations in the EM approach. Finally, we note that iCluster+ is not feasible in scenario 3 because it does not support binomial variables, while GBFA_CL works seamlessly.

V. Real Data Analysis

A. NCI60 Cell Line Data

The NCI60 is a panel of 60 diverse human cancer cell lines used by the Developmental Therapeutics Program (DTP) of the U.S. National Cancer Institute (NCI) to screen over 100,000 chemical compounds and natural products. It consists of 9 kinds of cancerous cell lines; leukemia, melanomas, ovarian, renal, breast, prostate, colon, lung, and CNs origin. There are various -omics datasets available for those cell lines including gene expression data from various platforms, protein abundance data, and methylation data. We download three datasets from CellMiner [35], two of which are gene expression data and the other one is protein abundance data.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

The first data is a transcript profile data based on Affymetrix HG-U133 chips [36]. The second data is a mRNA expression data based on Agilent Whole Human Genome Oligo Microarray technology [37]. The last one is a proteomics profiling data using high-density reverse-phase lysate microarrays [38]. We use 59 cell line data consisting of 9 subgroups which are common to all three datasets. As a preprocessing [39], we select the top 5% of genes with high variance, which results in 491 genes in the affymetrix data, 488 genes in the agilent data, and 94 proteins in proteomics data.

We incorporate the pathway graph information from the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database [40] in GBFA_CL. For the pathway enrichment data analysis, we use ToppGene Suite [41], a portal for the enrichment analysis.

B. Results

We apply GBFA_CL and iCluster+ to NCI60 data. Various combinations of tuning parameters are explored for both methods and the optimal model is selected in the same manner as in section IV.

We compare the estimated clusterings with the true cell line clustering. The results are shown in Table V. iCluster+ chooses $L = 8$ (and $K = 9$) and GBFA_CL chooses similarly between $L = 8, 9$ and $K = 8, 9, 10$. In terms of the similarity of the clustering, all versions of GBFA_CL show better results except for the case with $\eta = 0.5$, in which case the result is comparable to that of iCluster+. Overall, the incorporation of the pathway graph knowledge improves the similarity of the estimated clustering to the truth.

We also conduct the pathway enrichment analysis to identify the pathways significantly enriched among the genes selected by GBFA factor loadings. We compare the genes selected from GBFA_CL with $\eta = 0$ and $\eta = 1.5$, and the pathways that are related with cancers are listed in in Table IV. “Pathways in cancer” is enriched in most of the factors (factor loadings). Other cancer related pathways such as breast cancer, melanoma, lung cancer, and prostate cancer are also enriched in several factors. “Hemostatis” and “Melanogenesisis” are also related to leukemia [42] and melanoma [43, 44] cancers, respectively. These facts indicate that GBFA successfully captures the underlying biological variations.

We observe that there are two pathways enriched only in GBFA_CL with graph information incorporated; the neurotrophin signaling pathway and the chemokine signaling pathway. It is known that the neurotrophins regulate the cancer stem cells [45] and the chemokine signaling pathway governs cancer progression [46]. Figure 2 shows the edges between the genes in our data which belong to the neurotrophin signaling pathway. Incorporating this graph information has lead to the conclusion that can potentially be biologically meaningful.

VI. Conclusion and Discussion

We have proposed the generalized Bayesian factor analysis framework, which is a generalization of the sparse factor analysis framework in two ways; the incorporation of the network graph information and the accommodation of multiple data types. We have also proposed the very efficient variational EM algorithm for the MAP estimator. Then, GBFA

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

has been applied to the integrative clustering analysis. The newly proposed integrative clustering method has been proven to be flexible, accurate, and efficient throughout the simulation studies and the application to the NCI60 datasets. Moreover, we have shown that the use of the proposed method on the analysis of multi-omics data with incorporation of the pathway graph information can deliver more biologically meaningful outcomes.

Note that the applications of our GBFA framework is not limited to the integrative clustering only. We believe that GBFA can extend many other existing statistical methods and can potentially address plenty of problems where the structural graph information exists.

Another future work will include the extension of the proposed framework for more complicated forms of data. Due to the advances of the mobile businesses and technologies, various biomedical data such as the electronic health record (EHR) are available in great abundance. Such data often involve not only graphical structure but also hierarchical structure, multi class categorical variables, missing values, and distributed repositories.

Acknowledgement

This work is partly supported by NIH grants P30CA016520, R01DK108070, and R01GM124111. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1]. Wang D and Gu J, “Integrative clustering methods of multi-omics data for molecule-based cancer classifications,” Quantitative Biology, vol. 4, no. 1, pp. 58–67, 2016.

[2]. Kristensen VN, Lingjærde OC, Russnes HG, Vollan HKM, Frigessi A, and Børresen-Dale A-L, “Principles and methods of integrative genomic analyses in cancer,” Nature Reviews Cancer, vol. 14, no. 5, pp. 299–313, 2014. [PubMed: 24759209]

[3]. Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y et al., “The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups,” Nature, vol. 486, no. 7403, p. 346, 2012. [PubMed: 22522925]

[4]. Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, Powers RS, Ladanyi M, and Shen R, “Pattern discovery and cancer gene identification in integrated cancer genomic data,” Proceedings of the National Academy of Sciences, vol. 110, no. 11, pp. 4245–4250, 2013.

[5]. Chalise P and Fridley BL, “Integrative clustering of multi-level omic data based on non-negative matrix factorization algorithm,” PloS one, vol. 12, no. 5, p. e0176278, 2017. [PubMed: 28459819]

[6]. Shen R, Olshen AB, and Ladanyi M, “Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis,” Bioinformatics, vol. 25, no. 22, pp. 2906–2912, 2009. [PubMed: 19759197]

[7]. Shen R, Wang S, and Mo Q, “Sparse Integrative Clustering of Multiple Omics Data Sets,” Annals of Applied Statistics, vol. 7, no. 1, pp. 269–294, 2013. [PubMed: 24587839]

[8]. Kim S, Oesterreich S, Kim S, Park Y, and Tseng GC, “Integrative clustering of multi-level omics data for disease subtype discovery using sequential double regularization,” Biostatistics, vol. 18, no. 1, pp. 165–179, 2017. [PubMed: 27549122]

[9]. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, Natarajan KN, Reik W, Barahona M, Green AR, and Hemberg M, “SC3: consensus clustering of single-cell RNA-seq data,” Nature Methods, vol. 14, no. 5, pp. 483–486, 3 2017. [PubMed: 28346451]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

[10]. Mo Q, Shen R, Guo C, Vannucci M, Chan KS, and Hilsenbeck SG, “A fully bayesian latent variable model for integrative clustering analysis of multi-type omics data,” Biostatistics, vol. 19, no. 1, pp. 71–86, 2018. [PubMed: 28541380]

[11]. S. C. F., “Pathway databases,” Annals of the New York Academy of Sciences, vol. 1020, no. 1, pp. 77–91.

[12]. Kanehisa M, Furumichi M, Tanabe M, Sato Y, and Morishima K, “Kegg: new perspectives on genomes, pathways, diseases and drugs,” Nucleic Acids Research, vol. 45, no. D1, pp. D353–D361, 2017. [PubMed: 27899662]

[13]. Li C and Li H, “Network-constrained regularization and variable selection for analysis of genomic data,” Bioinformatics, vol. 24, no. 9, pp. 1175–1182, 2008. [PubMed: 18310618]

[14]. Pan W, Xie B, and Shen X, “Incorporating predictor network in penalized regression with application to microarray data,” Biometrics, vol. 66, no. 2, pp. 474–484, 2010. [PubMed: 19645699]

[15]. Li F and Zhang NR, “Bayesian Variable Selection in Structured High-Dimensional Covariate Spaces with Applications in Genomics,” Journal of the American Statistical Association, vol. 105, no. 491, pp. 1202–1214, 2010.

[16]. Stingo FC, Chen YA, Tadesse MG, and Vannucci M, “Incorporating Biological Information into Linear Models: A Bayesian Approach to the Selection of Pathways and Genes,” Annals of Applied Statistics, vol. 5, no. 3, pp. 1978–2002, 2011. [PubMed: 23667412]

[17]. Chang C, Kundu S, and Long Q, “Scalable bayesian variable selection for structured highdimensional data,” Biometrics, vol. 0, no. 0.

[18]. Rockova V and Lesaffre E, “Incorporating Grouping Information in Bayesian Variable Selection with Applications in Genomics,” Bayesian Analysis, vol. 9, no. 1, pp. 221–258, 2014.

[19]. Yu G and Liu Y, “Sparse regression incorporating graphical structure among predictors,” Journal of the American Statistical Association, vol. 111, no. 514, pp. 707–720, 2016. [PubMed: 29503486]

[20]. Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, Haussler D, and Stuart JM, “Inference of patient-specific pathway activities from multidimensional cancer genomics data using paradigm,” Bioinformatics, vol. 26, no. 12, pp. i237–i245, 2010. [PubMed: 20529912]

[21]. Klami A, Virtanen S, Leppaaho E, and Kaski S, “Group factor analysis,” Neural Networks and Learning Systems, IEEE Transactions on, vol. 26, no. 9, pp. 2136–2147, 2015.

[22]. Virtanen S, Klami A, Khan SA, and Kaski S, “Bayesian group factor analysis,” in Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS-12), vol. 22, 2012, pp. 1269–1277.

[23]. Polson NG, Scott JG, and Windle J, “Bayesian inference for logistic models using pólya-gamma latent variables,” Journal of the American statistical Association, vol. 108, no. 504, pp. 1339–1349, 2013.

[24]. Ročková V and George EI, “Fast Bayesian Factor Analysis via Automatic Rotations to Sparsity,” Journal of the American Statistical Association, vol. 111, no. 516,pp. 1608–1622, 2016.

[25]. Ročková V and George EI, “The spike-and-slab lasso,” Journal of the American Statistical Association, vol. 113, no. 521, pp. 431–444, 2018.

[26]. George EI and Mcculloch RE, “Variable Selection via Gibbs Sampling,” Journal of the American Statistical Association, vol. 88, no. 423, pp. 881–889, 1993.

[27]. Blei DM, Kucukelbir A, and McAuliffe JD, “Variational inference: A review for statisticians,” Journal of the American Statistical Association, vol. 112, no. 518, pp. 859–877, 2017.

[28]. Becker S and Le Cun Y, “Improving the convergence of back-propagation learning with second order methods,” in Proceedings of the 1988 connectionist models summer school, 1988, pp. 29–37.

[29]. Chang C and Tsay RS, “Estimation of Covariance Matrix via the Sparse Cholesky Factor with Lasso,” Journal of Statistical Planning and Inference, vol. 140, pp. 3858–3873, 2010.

[30]. The Elbow Method, “The elbow method — Wikipedia, the free encyclopedia,” 2018 [Online]. Available: https://en.wikipedia.org/wiki/Elbow_method_(clustering)

[31]. Rousseeuw PJ, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” Journal of computational and applied mathematics, vol. 20, pp. 53–65, 1987.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

[32]. Tibshirani R, Walther G, and Hastie T, "Estimating the number of clusters in a data set via the gap statistic," Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 63, no. 2, pp. 411–423, 2001.

[33]. Jaccard P, "Nouvelles recherches sur la distribution floral," Bull. Soc. Vard. Sci. Nat, vol. 44, pp. 223–270, 1908.

[34]. Rand WM, "Objective criteria for the evaluation of clustering methods," Journal of the American Statistical association, vol. 66, no. 336, pp. 846–850, 1971.

[35]. Reinhold WC, Sunshine M, Liu H, Varma S, Kohn KW, Morris J, Doroshow J, and Pommier Y, "Cellminer: A web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the nci-60 cell line set," Cancer Research, vol. 72, no. 14, pp. 3499–3511, 2012. [PubMed: 22802077]

[36]. Shankavaram UT, Reinhold WC, Nishizuka S, Major S, Morita D, Chary KK, Reimers MA, Scherf U, Kahn A, Dolginow D et al., "Transcript and protein expression profiles of the nci-60 cancer cell panel: an integromic microarray study," Molecular cancer therapeutics, vol. 6, no. 3, pp. 820–832, 2007. [PubMed: 17339364]

[37]. Liu H, D'Andrade P, Fulmer-Smentek S, Lorenzi P, Kohn KW, Weinstein JN, Pommier Y, and Reinhold WC, "mrna and microrna expression profiles of the nci-60 integrated with drug activities," Molecular cancer therapeutics, vol. 9, no. 5, pp. 1080–1091, 2010. [PubMed: 20442302]

[38]. Nishizuka S, Charboneau L, Young L, Major S, Reinhold WC, Waltham M, Kouros-Mehr H, Bussey KJ, Lee JK, Espina V et al., "Proteomic profiling of the nci-60 cancer cell lines using new high-density reversephase lysate microarrays," Proceedings of the National Academy of Sciences, vol. 100, no. 24, pp. 14 229–14234, 2003.

[39]. Cai T and Liu W, "A direct estimation approach to sparse linear discriminant analysis," Journal of the American Statistical Association, vol. 106, no. 496, pp. 1566–1577, 2011.

[40]. Kanehisa M, Furumichi M, Tanabe M, Sato Y, and Morishima K, "KEGG: new perspectives on genomes, pathways, diseases and drugs," Nucleic acids research, vol. 45, no. D1, pp. D353–D361, 2016. [PubMed: 27899662]

[41]. Chen J, Bardes EE, Aronow BJ, and Jegga AG, "Toppgene suite for gene list enrichment analysis and candidate gene prioritization," Nucleic acids research, vol. 37, no. suppl_2, pp. W305–W311, 2009. [PubMed: 19465376]

[42]. Lewis JH, Burchenal JH, Ellison RR, Ferguson JH, Palmer JH, Murphy ML, Zucker MB, Morgan F, and Rudin I, "Studies of hemostatic mechanisms in leukemia and thrombocytopenia," American journal of clinical pathology, vol. 28, no. 5, pp. 433–446, 2016.

[43]. Slominski A, Kim T-K, Brożyna A, Janjetovic Z, Brooks D, Schwab L, Skobowiat C, Jóźwicki W, and Seagroves T, "The role of melanogenesis in regulation of melanoma behavior: Melanogenesis leads to stimulation of hif-1α expression and hif-dependent attendant pathways," Archives of biochemistry and biophysics, vol. 563, pp. 79–93, 2014. [PubMed: 24997364]

[44]. Meira WV, Heinrich TA, Cadena SMSC, and Martinez GR, "Melanogenesis inhibits respiration in b16-f10 melanoma cells whereas enhances mitochondrial cell content," Experimental Cell Research, vol. 350, no. 1, pp. 62–72, 2017. [PubMed: 27864061]

[45]. Chopin V, Lagadec C, Toillon R-A, and Le Bourhis X, "Neurotrophin signaling in cancer stem cells," Cellular and molecular life sciences, vol. 73, no. 9, pp. 1859–1870, 2016. [PubMed: 26883804]

[46]. Hembruff SL and Cheng N, "Chemokine signaling in cancer: Implications on the tumor microenvironment and therapeutic targeting," Cancer therapy, vol. 7, no. A, p. 254, 2009. [PubMed: 20651940]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

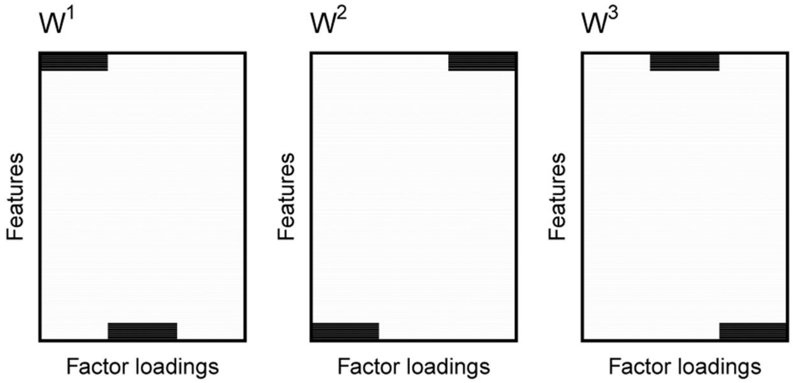

Fig. 1.

Regions of nonzero elements in W^1, W^2, and W^3 shown in black.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

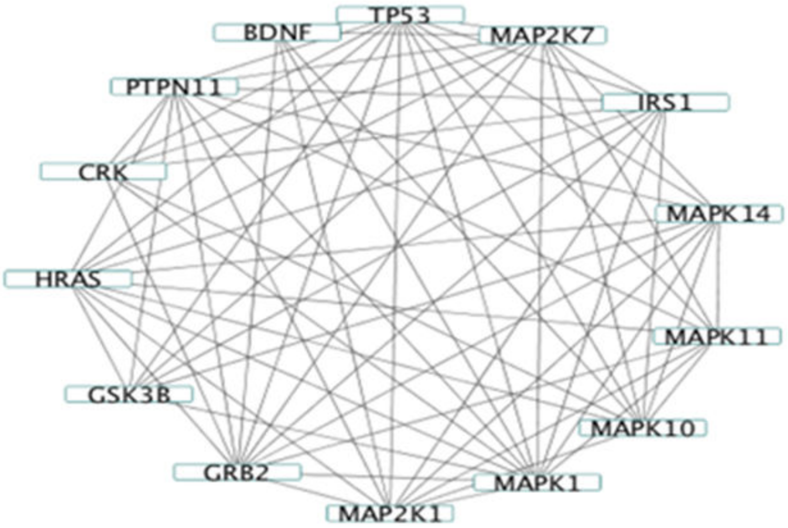

Fig. 2.

Network of Neurotrophin signaling pathway. Each node are genes selected by GBFA incorporating the network information.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

TABLE I

Parameters for unified likelihood in (6).

Type	$\boldsymbol{\psi_{ji}}$	$\boldsymbol{k_{ji}}$	$\pi_j^*(\boldsymbol{\rho}_j)$
Gaussian	x_{ji}	0	$\rho_{ji} \equiv \rho_j \sim \mathcal{G}\left(\frac{\zeta_j + n}{2}, \frac{\zeta_j}{2}\right)$
Binomial	0	$x_{ji} - n_j/2$	$\rho_{ji} \sim \mathcal{PG}(n_j, 0)$
Neg. Bin.	0	$(x_{ji} - r_j)/2$	$\rho_{ji} \sim \mathcal{PG}(x_{ji} + r_j, 0)$

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

TABLE II

Combinations of data types explored in each simulation scenario.

Scenario	X^1	X^2	X^3
1	Gaussian	Gaussian	Gaussian
2	Gaussian	Bernoulli	Poisson
3	Gaussian	Binomial	Poisson

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

TABLE III

Summary of the performance in the simulation study. The competing methods are the *K*-means clustering applied to the concatenated data and separated data, iCluster+, the integrative clustering proposed by Mo et al. [4]; and GBFA_CL, the proposed method. Rate indicates the rate of choosing the right number of clusters (*K*). Averages of Jaccard and Rand indexes out of 100 Monte Carlo datasets are reported. The corresponding standard errors are shown in the parentheses.

$K = 4$										
		Scenario 1			Scenario 2			Scenario 3		
Method		Rate	Jaccard	Rand	Rate	Jaccard	Rand	Rate	Jaccard	Rand
con *K*-means		0.52	0.441(.011)	0.763(.010)	0.19	0.322(.007)	0.606(.014)	0.19	0.308(.006)	0.559(.014)
	$H = 1$	0.04	0.289(.005)	0.622(.007)	0.03	0.288(.004)	0.623(.006)	0.02	0.291(.004)	0.620(.007)
sep *K*-means	$H = 2$	0.05	0.287(.004)	0.618(.006)	0.01	0.408(.007)	0.715(.008)	0.00	0.550(.008)	0.811(.007)
	$H = 3$	0.03	0.283(.003)	0.599(.004)	0.16	0.318(.007)	0.606(.014)	0.21	0.311(.006)	0.583(.015)
iCluster+		0.17	0.321(.007)	0.674(.008)	0.37	0.603(.012)	0.877(.006)		-	-
	$\eta = 0$	0.91	0.648(.001)	0.893(.000)	0.76	0.704(.002)	0.880(.001)	0.77	0.792(.001)	0.931(.001)
GBFA_CL	$\eta = 0.5$	0.97	0.675(.001)	0.903(.000)	0.77	0.706(.002)	0.879(.001)	0.82	0.807(.001)	0.938(.001)
	$\eta =1$	0.98	0.684(.001)	0.906(.000)	0.78	0.709(.002)	0.883(.001)	0.84	0.810(.001)	0.940(.001)

$K = 6$										
		Scenario 1			Scenario 2			Scenario 3		
Method		Rate	Jaccard	Rand	Rate	Jaccard	Rand	Rate	Jaccard	Rand
con *K*-means		0.01	0.261(.003)	0.618(.004)	0.02	0.288(.006)	0.635(.014)	0.02	0.303(.007)	0.663(.014)
	H=1	0.00	0.218(.001)	0.581(.003)	0.00	0.217(.002)	0.579(.003)	0.00	0.215(.001)	0.570(.002)
sep *K*-means	H=2	0.00	0.212(.002)	0.570(.002)	0.13	0.174(.003)	0.643(.010)	0.00	0.238(.002)	0.592(.002)
	H=3	0.00	0.214(.002)	0.578(.003)	0.00	0.298(.006)	0.659(.014)	0.02	0.306(.007)	0.670(.014)
iCluster+		0.28	0.341(.010)	0.825(.004)	0.25	0.508(.006)	0.825(.001)		-	-
	$\eta = 0$	0.85	0.580(.012)	0.907(.004)	0.62	0.647(.020)	0.897(.010)	0.91	0.811(.011)	0.963(.003)
GBFA_CL	$\eta = 0.5$	0.89	0.631(.010)	0.922(.003)	0.69	0.687(.017)	0.919(.009)	0.87	0.809(.011)	0.962(.003)
	$\eta =1$	0.92	0.650(.011)	0.927(.003)	0.63	0.667(.017)	0.913(.009)	0.85	0.798(.012)	0.960(.003)

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

TABLE IV

Pathway enrichment analysis result for NCI60 data analysis results of the GBFA. Source means the database that has information about corresponding pathway, and *q*-value is the Bonferroni adjusted *p*-value. *L* is the index of the factor loading that includes the genes of the enriched pathway. REACTOME is another biological pathway database like KEGG pathway database.

Name	Source	$\eta = 0$		$\eta = 1.5$	
		L	*q*-value	*L*	*q*-value
Pathways in cancer	REACTOME	2	8.11*e*–10	2	1.27*e*–09
Non-small cell lung cancer	KEGG	4	1.09*e*–05	4	1.16*e*–05
Bladder cancer	KEGG	4	8.46*e*–05	4	8.88*e*–05
Melanogenesis	KEGG	4	1.77*e*–04	2	4.15*e*–04
Breast cancer	KEGG	2	1.13*e*–03	2	1.36*e*–03
Chronic myeloid leukemia	KEGG	4	8.37*e*–03	4	8.78*e*–03
Prostate cancer	KEGG	6	1.42*e*–02	6	1.62*e*–02
Hemostasis	REACTOME	1	3.71*e*–02	1	2.57*e*–02
Neurotrophin signaling pathway	KEGG	-	-	6	2.77*e*–02
Chemokine signaling pathway	KEGG	-	-	6	3.23*e*–02

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

TABLE V

NCI60 data analysis results from iCluster+ and 4 different versions of GBFA_CL ($\eta = 0, 0.5, 1, 1.5$). F*or each method, the chosen number of latent factors and the chosen number of clusters are listed*. Two similarity measures are also listed.

Method	η	L	K	**Jaccard**	**Rand**
iCluster+	-	8	9	0.348	0.896
GBFA_CL	0	9	10	0.496	0.930
	0.5	8	8	0.367	0.886
	1	9	9	0.521	0.933
	1.5	9	10	0.527	0.933