Generalized Bayesian Factor Analysis Framework

Jingxuan Bao

Advisor: Dr. Qi Long, Dr. Changgee Chang
Dr. Shen's Lab Group Meeting

School of Arts and Sciences
Applied Math and Computational Science
University of Pennsylvania

September 11, 2020

Overview

1. Introduction
2. Model Formulation
3. Unify Likelihood Functions
4. Variational EM Algorithm

Motivation

Motivation

What is a Generalized Bayesian Factor Analysis (GBFA) model?

The GBFA is used for extraction of the common factors.

Why we need GBFA model?

Example: Clustering - Grouping about the data.

What is Factor Analysis Model?

Suppose we have a set of p observable random variables $x_1, \dots, x_p$, with means $\mu_1, \dots, \mu_p$. Suppose for some unknown constants l_{ij} and unobserved random variables F_j , where $i \in 1, 2, \dots, p$ and $j \in 1, 2, \dots, k$, where $k < p$, we have that the terms in each random variable should be writeable as a linear combination of the common factors:

$$x_i - \mu_i = l_{i1} F_1 + \cdots + l_{ik} F_k + \epsilon_i.$$

$$\begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \\ \dots \\ x_p - \mu_p \end{bmatrix} = \begin{bmatrix} l_{11} & l_{12} & \dots & l_{1k} \\ l_{21} & l_{22} & \dots & l_{2k} \\ \dots & \dots & \dots & \dots \\ l_{p1} & l_{p2} & \dots & l_{pk} \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \\ \dots \\ F_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \dots \\ \epsilon_p \end{bmatrix}$$

Here, the ϵ_i are unobserved stochastic error terms with zero mean and finite variance, which may not be the same for all i.

Model Formulation

Model Formulation

Suppose we have H multi-modal data generated from various technologies, say $X^1_{p_1 \times n}, \dots, X^H_{p_H \times n}$. Denote

$$X = \begin{bmatrix} X^1_{p_1 \times n} \\ X^2_{p_2 \times n} \\ \dots \\ X^H_{p_H \times n} \end{bmatrix} \in \mathbb{R}^{p \times n}$$

where $p = \sum_{h=1}^{H} p_h$.

Model Formulation

Example: Three datasets from CellMiner, two of which are gene expression data and the other one is protein abundance data.

- The first data - a transcript profile data based on Affymetrix HG-U133 chips;
- The second data - a mRNA expression data based on Agilent Whole Human Genome Oligo Microarray technology;
- The last data - a proteomics profiling data using high-density reverse-phase lysate microarrays.

We use 59 cell line data consisting of 9 subgroups which are common to all three datasets. 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.

Model Formulation

Suppose x_{ji} for all $1 \leq i \leq n$ belong to the same distribution family and the distribution of X has the following form of likelihood which is governed by the parameter matrix μ:

$$\pi(X|\boldsymbol{\mu}) = \prod_{j=1}^{p} \underbrace{\prod_{i=1}^{n} \pi_j(x_{ji}|\mu_{ji})}_{the\ j-th\ row}$$

where in this model, we can assume x_{ji} follows the binomial, negative binomial, Poisson, and Gaussian distributions.

What is $\boldsymbol{\mu}$?

We assume Generalized Bayesian Factor Analysis Model:

Let $\boldsymbol{z}_i$ be the L-dimensional latent factor for the i-th subject and $\tilde{\boldsymbol{w}}_j$ be the factor loadings for the j-th feature. Let

$$\begin{aligned}\boldsymbol{\mu} &= \boldsymbol{m}\mathbf{1}^T + WZ \\ &= \boldsymbol{m}\mathbf{1}^T + \begin{bmatrix}\tilde{\boldsymbol{w}}_1 \\ \dots \\ \tilde{\boldsymbol{w}}_p\end{bmatrix}_{p\times L} \begin{bmatrix}\boldsymbol{z}_1 & \dots & \boldsymbol{z}_n\end{bmatrix}_{L\times n}\end{aligned}$$

where $L \ll p$, $\boldsymbol{m}$ is a random variable which represents the location of the data.

Model Formulation

Recall that we assume

$$\pi(X|\boldsymbol{\mu}) = \prod_{j=1}^{p} \underbrace{\prod_{i=1}^{n} \pi_j(x_{ji}|\mu_{ji})}_{the\ j-th\ row}$$

where in this model, we can assume x_{ji} follows the binomial, negative binomial, Poisson, and Gaussian distributions.

Model Formulation

Binomial distribution:

$$\pi_j(x_{ji}|\mu_{ji}) = \binom{n_j}{x_{ji}} \frac{e^{\mu_{ji}x_{ji}}}{(1+e^{\mu_{ji}})^{r_j+x_{ji}}}, \quad 0 \leq x_{ji} \leq n_j$$

Negative Binomial distribution:

$$\pi_j(x_{ji}|\mu_{ji}) = \binom{r_j + x_{ji} - 1}{x_{ji}} \frac{e^{\mu_{ji}x_{ji}}}{(1+e^{\mu_{ji}})^{r_j+x_{ji}}}, \quad x_{ji} \geq 0$$

Poisson distribution:

$$\pi_j(x_{ji}|\mu_{ji}) = e^{-e^{\mu_{ji}}} \frac{e^{\mu_{ji}x_{ji}}}{x_{ji}!}, \quad x_{ji \geq 0}$$

Gaussian distributions:

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

Prior for Z

We consider the standard Gaussian prior for Z since the standard orthonormality assumption on the latent factors,

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

Prior for W

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

$$\log \pi(W|\gamma) = C + \sum_{j=1}^{p}\sum_{l=1}^{L} \log \lambda_{jl} - \sum_{j=1}^{p}\sum_{l=1}^{L} \lambda_{jl}|w_{jl}|$$

where $\lambda_{jl} = (1-\gamma_{jl})\lambda_0 + \gamma_{jl}\lambda_1$, $0 \leq \lambda_1 \leq \lambda_0$.

Intuition of Spike and Slab Prior

We assume spike and slab prior to incorporate the graph information.

Intuition of Spike and Slab Prior

What we want is group-wised selection, i.e., if two genes are correlated, we want to select them both. So, if two genes are correlated, they will at least share one common factor. We encourage the adjacent variables (correlated genes) in our prior graph information to load the same factors.
Specifically, we assume the mean of data $x_i, i \in \{1, \ldots, p\}$ is a linear combination of factors $z_j, j \in \{1, \ldots, k\}$ plus a location variable.

$$\mu_{ij} = m_i + \omega_{i1} z_{i1} + \omega_{i2} z_{i2} + \cdots + \omega_{iL} z_{iL}.$$

If x_j and x_k are adjacent in $\mathcal{G}$ and if $\omega_{jl} \neq 0$ for some l, then we promote $w_{kl} \neq 0$.

Incorporating Graph Information

We consider to incorporating network information into the prior for γ.

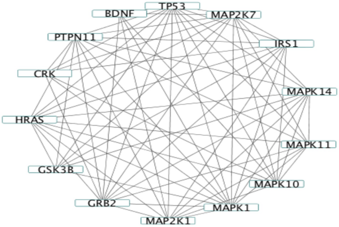

Figure: Example of Graph Information

Suppose the graphs G is the adjacency matrix for graph $\langle P, E\rangle$, which is obtained by combining the graphs $\langle P_h, E_h\rangle$, where the presence of edge indicates the correlation between the relevant pair of variables.

According to this set up, we achieve that 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.

Incorporating Graph Information - Prior for γ

Therefore, it is reasonable to encourage the pairs of adjacent variables to share common factors.

We can employ the Markov random field prior for γ,

$$\log \pi(\gamma) = C_{\delta,\eta} - \delta \sum_{j=1}^{p} \sum_{l=1}^{L} \gamma_{jl} + \eta \sum_{j=1}^{p} \sum_{k=1}^{p} \sum_{l=1}^{L} G_{jk} \gamma_{jl} \gamma_{kl}.$$

We can also empoly the Ising prior for γ,

$$\log \pi(\gamma) = C_{\delta,\eta} - \delta \sum_{j=1}^{p} \sum_{l=1}^{L} \gamma_{jl} + \eta \sum_{j=1}^{p} \sum_{k=1}^{p} \sum_{l=1}^{L} G_{jk} \mathbb{I}(\gamma_{jl} = \gamma_{kl}).$$

Model Formulation - Summary

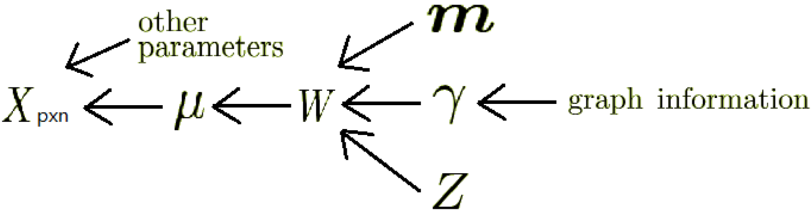

Figure: Summary of Model Formulation

Unify Likelihood Functions

Unify Likelihood Functions

Recall that we assume

$$\pi(X|\boldsymbol{\mu}) = \prod_{j=1}^{p} \underbrace{\prod_{i=1}^{n} \pi_j(x_{ji}|\mu_{ji})}_{the\ j-th\ row}$$

where in this model, we can assume x_{ji} follows the binomial, negative binomial, Poisson, and Gaussian distributions.

Now, we want to unify the likelihood functions for different distributions. We present the identity

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

where $\kappa = x_{ji} - \frac{b_{ji}}{2}$ and $\pi(\rho_{ji}) = \mathcal{PG}(b_{ji}, 0)$.

Unify Likelihood Functions

Now we apply the identity to the binomial, negative binomial, and Poisson likelihood functions, then we will obtain

$$\pi_j(\tilde{\boldsymbol{x}}_j, \tilde{\boldsymbol{\rho}}_j|\tilde{\boldsymbol{\mu}}_j) = \pi_j(\tilde{\boldsymbol{x}}_j|\tilde{\boldsymbol{\mu}}_j)\pi^*(\tilde{\boldsymbol{\rho}}_j) \propto e^{-\frac{1}{2}\sum_i \rho_{ji}(\mu_{ji}-\psi_{ji})^2+\sum_i \kappa_{ji}\mu_{ji}}\pi_j^*(\tilde{\boldsymbol{\rho}}_j)$$

where the unknown parameters are shown in Table 1.

Type	ψ_{ji}	κ_{ji}	$\pi_j^*(\boldsymbol{\rho}_j)$
Gaussian	x_{ji}	0	$\rho_{ji} = \boldsymbol{\rho}_j \sim \mathcal{G}(\frac{\zeta_j+n}{2}, \frac{\zeta_j}{2})$
Binomial	0	$x_{ji} - \frac{n_j}{2}$	$\rho_{ji} \sim \mathcal{PG}(n_j, 0)$
Negative Binomial	0	$\frac{x_{ji}-r_j}{2}$	$\rho_{ji} \sim \mathcal{PG}(x_{ji} + r_j, 0)$

Table: Parameters for Unified Likelihood

Unify Likelihood Functions

Now, the full likelihood can be written as

$$\pi(\gamma, \boldsymbol{\rho}, W, Z, X|\boldsymbol{m}) = \pi(W|\gamma)\pi(\gamma)\pi(Z)\prod_j \pi_j(\tilde{\boldsymbol{x}}_j|\tilde{\boldsymbol{\mu}}_j)\pi_j^*(\tilde{\boldsymbol{\rho}}_j).$$

Variational EM Algorithm

Maximum a Posteriori Estimation and EM Algorithm

In Bayesian statistics, a maximum a posteriori probability (MAP) estimate is an estimate of an unknown quantity, that equals the mode of the posterior distribution.

When dealing with the problem containing latent variable (γ), we often need to first marginalize the latent variable out and then take derivation.

Problems

- Somehow it is impossible to marginalize γ but calculate the expected value of log posterior likehood is somewhat doable.
- Even if the marginalization of γ is possible, the calculation is not easy.

Solution

EM Algorithm

EM Algorithm and Variational EM Algorithm

When doing E-step, sometimes the classical EM approach involves an intractable conditional expectation of the log-likelihood.

To address the problem, we use the latent variable augmentation technique and the variational EM approach.

Variational EM Algorithm

Since the EM Algorithm does not have analytic solution to the conditional expectations involving γ, ρ, Z given W. We consider the variation EM approach. We consider a product measure on individual γ_{jl}, ρ, Z and let $\hat{\pi}(\gamma, \rho, Z) = \hat{\pi}(\gamma)\hat{\pi}(\rho)\hat{\pi}(Z)$ where

$$\hat{\pi}(\gamma) = \prod_j \prod_l \theta_{jl}^{\gamma_{jl}} (1 - \theta_{jl})^{1-\gamma_{jl}}$$

$$\hat{\pi}(\rho) \propto \prod_j e^{-\frac{1}{2}\sum_i \rho_{ji}\varphi_{ji}^2} \pi_j^*(\tilde{\boldsymbol{\rho}}_j)$$

$$\hat{\pi}(Z) \propto \prod_i e^{-\frac{1}{2}(\boldsymbol{z}_i - \boldsymbol{\mu}_{z,i})^T \Sigma_{z,i}^{-1} (\boldsymbol{z}_i - \boldsymbol{\mu}_{z,i})}$$

where $\hat{\mathbb{E}}$ is the expectation operator under $\hat{\pi}$. Note that $\boldsymbol{\theta}$, $\boldsymbol{\varphi}$, $\boldsymbol{\mu}$ and σ are variational parameter which are not fixed yet.

Variational EM Algorithm - Kullback Leibler Divergence

Problem

The variational measure is not equal to the actual conditional distribution, which cause the calculation of expectation not precise.

Solution

We need to make the variational measure as closed as the actual conditional distribution. It is natural to consider a distance which measure two different distribution.

Kullback Leibler Divergence

Variational EM Algorithm - Kullback Leibler Divergence

The Kullback–Leibler divergence from Q to P is defined to be

$$D_{KL}(P||Q) = -\mathbb{E}_P \log Q + \mathbb{E}_P \log Q.$$

Here, we let P to be variational measure and Q to be actual conditional distribution.

Here the smaller the KL divergence, the "closer" the two distributions.

To make the variational measure as close to the actual conditional distribution as possible, we need to minimize the KL divergence between the variational measure and actual conditional distribution.

Variational EM Algorithm - Variational Parameter

Minimize

$$D_{KL}(P||Q) = -\mathbb{E}_P \log Q + \mathbb{E}_P \log Q$$

with respect to the variational parameters. We will obtain the update formula for variational parameters.

Variational EM Algorithm - Variational EM Algorithm

E-Step:

Take expectation to the log posterior distribution with respect to latent variables using variational measure.

M-Step:

Find the optimizer to maximize the formula we obtained in E-Step to obtain the update formula for the other model parameters.

Variational EM Algorithm - Variational EM Algorithm

Algorithm 1

- Initialization;
- Iteration:
  - Update variational parameters until convergence;
  - Iteration:
    - Calculate E-Step using obtained variational measure;
    - Calculate M-Step to update the other model parameters until convergence;

Algorithm 2

- Initialization;
- Iteration until convergence:
  - Update variational parameters once;
  - Iteration:
    - Calculate E-Step using obtained variational measure;
    - Calculate M-Step to update the other model parameters once;

Variational EM Algorithm - ELBO

We can merge the step of minimization of KL divergence and the E-step and M-step together by using ELBO formula.

The definition of ELBO formula:

$$ELBO(P) = \mathbb{E}_P \log Q(\boldsymbol{z}, \boldsymbol{x}) - \mathbb{E}_P \log P(\boldsymbol{z})$$

where Q is posterior density and P is variational measure; and $\boldsymbol{z}$ is the set of variational parameters, $\boldsymbol{x}$ is the set of model parameters. The expectation is taken with respect to the P measure.

After we calculated the ELBO,

- we maximize it with respect to the variational measure parameters;
- we maximize the formula with respect to the other model parameters;

and we just repeat the step until convergence.

Variational EM Algorithm - Intuition of ELBO

Recall the KL divergence:

$$D_{KL}(P||Q) = -\mathbb{E}_P \log Q(\boldsymbol{z}, \boldsymbol{x}) + \mathbb{E}_P \log P(\boldsymbol{z})$$

Recall the ELBO formula:

$$ELBO(P) = \mathbb{E}_P \log Q(\boldsymbol{z}, \boldsymbol{x}) - \mathbb{E}_P \log P(\boldsymbol{z})$$

Thank You