Phenotyping via Bayesian Nonparametric Tensor Factorization

***Abstract*—We do phenotyping.**

I. Introduction

II. Model

A. CANDECOMP/PARAFAC (CP) Tensor Decomposition

A tensor $T \in R^{n_1 \times n_2 \times \ldots \times n_d}$ can be decomposed by a linear combination of rank-1 tensors:

$$T = \sum_{i=1}^{r} \lambda_i a_i^1 \otimes a_i^2 \ldots a_i^d$$

Where $\lambda_i \in R, a_i^k \in R^{n_k}$ and $\otimes$ represents the outer product. The r is the rank of tensor when r is minimized in above equation. When r is nor minimal, then the above decomposition is referred to as CANDECOMP/PARAFAC decomposition.

B. Dirichlet Process Tensor Factorization

TABLE I. Data Parameters

Notation	Meaning
λ	The weight vector of phenotypes, here $\|\lambda\| = 1$
λ_k	The weight of k-th phenotype; $\lambda_k \in [0, 1]$
N_p	Number of patients
N_m	Number of medications
N_d	Number of diagnosis
T	Record Tensor. $T \in \{0, 1\}^{N_p \times N_m \times N_d}$
N	Number of Non-zero elements in tensor T

TABLE II. Model Parameters

Notation	Meaning
α	The hyper parameter of Beta Distribution
β_i	The stick-breaking process random variable. $\beta_i \sim Beta(1, \alpha)$
$\gamma^{(p)}$	The hyper parameter of prior Dirichlet Distribution of patient
$\gamma^{(m)}$	The hyper parameter of prior Dirichlet Distribution of medication
$\gamma^{(d)}$	The hyper parameter of prior Dirichlet Distribution of diagnosis
$\theta^{(p,k)}$	The multinomial distribution over patients given the phenotype k
$\theta^{(m,k)}$	The multinomial distribution over medications given the phenotype k
$\theta^{(d,k)}$	The multinomial distribution over diagnosis given the phenotype k

We place a stick-breaking process prior on the λ. The parametere β_i follows the Beta Distribution $\beta_i \sim Beta(1, \alpha), i = 1, 2, \ldots$. And:

$$\lambda_k = \beta_k \prod_{i=1}^{k-1} (1 - \beta_i), k = 1, 2, \ldots$$

A $k-$dimensional Dirichlet random variable θ can take values in the $(k-1)$-simplex where $\sum_{i=1}^{k} \theta_i = 1$, and the probability density is

$$p(\theta|\gamma) = \frac{\mathcal{T}(\sum_{i=1}^{k} \gamma_i)}{\prod_{i=1}^{k} \mathcal{T}(\gamma_i)} \theta_1^{\gamma_1 - 1} \ldots \theta_k^{\gamma_k - 1}$$

It is natural to place the conjugate prior (Dirichlet Distribution) over the multinomial distributions parameterized by $\theta^{(p)}, \theta^{(m)}$ and $\theta^{(d)}$:

$$\begin{aligned}
\theta^{(p,k)} &\sim Dir(\gamma^{(p)}), k = 1, 2, \ldots \\
\theta^{(m,k)} &\sim Dir(\gamma^{(m)}), k = 1, 2, \ldots \\
\theta^{(d,k)} &\sim Dir(\gamma^{(d)}), k = 1, 2, \ldots
\end{aligned}$$

The patients, medications and diagnosis all follow the multinomial distribution parameterised by θ, specifically:

$$\begin{aligned}
\text{patient}_i &\sim Multi(\theta^{(p)}), i = 1, 2, \ldots, N_p \\
\text{medication}_i &\sim Multi(\theta^{(m)}), i = 1, 2, \ldots, N_m \\
\text{diagnosis}_i &\sim Multi(\theta^{(d)}), i = 1, 2, \ldots, N_d
\end{aligned}$$

Given the parameters, the likelihood of given record triplet (i, j, k) (patient i had diagnostic record k, and has taken medicine j) can be formulated as:

$$\begin{aligned}
&p(T_{i,j,k}|\alpha, \gamma^{(p)}, \gamma^{(m)}, \gamma^{(d)}) \\
&= \int p(\beta|\alpha) \sum_{r=1}^{\infty} \lambda_r p(i|r, \gamma^{(p)}) p(j|r, \gamma^{(m)}) p(k|r, \gamma^{(d)}) d\beta
\end{aligned}$$

C. Collapsed Gibbs Sampling

We define a set $S = \{(p_n, m_n, d_n)\}, n = 1, \ldots, N$, where $T(p_n, m_n, d_n) = 1$, consists of triplets representing the coordinates of non-zero elements in tensor T. Our model is a type of mixture model, so it is natural to define a hidden variable $z_n, n = 1, \ldots, N$ which representing the mixture index (i.e., the index of phenotype) of record triplet (p_n, m_n, d_n). z is a vector of length N and z_t represents its tth element.

Directly find the MAP of above equation is intractable, so we consider using MCMC to sample the parameters from posterior distribution. Here we use Gibbs sampling technique, then we need the posterior probability of a certain hidden variable z_t, $p(z_t|z_{-t}, S, \alpha, \gamma^{(p)}, \gamma^{(m)}, \gamma^{(d)})$. Here z_{-t} means the hidden variable set without tth element.

Let's consider a simpler case, when the number if phenotypes is R, a finite number. Then $\lambda \sim Dir(\frac{\alpha}{R})$. Use definition of conditional probability,

$$\begin{aligned}
p(z_t|z_{-t}, S, \alpha, \gamma^{(p)}, \gamma^{(m)}, \gamma^{(d)}) &= \frac{p(z_t, z_{-t}, S|\alpha, \gamma^{(p)}, \gamma^{(m)}, \gamma^{(d)})}{p(z_{-t}, S|\alpha, \gamma^{(p)}, \gamma^{(m)}, \gamma^{(d)})} \\
&\propto p(z_t, z_{-t}, S|\alpha, \gamma^{(p)}, \gamma^{(m)}, \gamma^{(d)}) \\
&= p(z, S|\alpha, \gamma^{(p)}, \gamma^{(m)}, \gamma^{(d)})
\end{aligned}$$

Expand it using rule of total probability,

$$
\begin{aligned}
&= \int\int\int\int p(z, S, \lambda|\theta^{(p)}, \theta^{(m)}, \theta^{(d)}) d\lambda d\theta^{(p)} d\theta^{(m)} d\theta^{(d)} \\
&= \int\int\int\int p(z|\lambda) p(\lambda|\alpha) p(S|\theta^{(p)}, \theta^{(m)}, \theta^{(d)}, z) \\
&\quad p(\theta^{(p)}|\gamma^{(p)}) p(\theta^{(m)}|\gamma^{(m)}) p(\theta^{(d)}|\gamma^{(d)}) d\lambda d\theta^{(p)} d\theta^{(m)} d\theta^{(d)} \\
&= \int p(z|\lambda) p(\lambda|\alpha) d\lambda \times \int\int\int \prod_{n=1}^{N} p(S_n|\theta^{(p)}, \theta^{(m)}, \theta^{(d)}, z) \\
&\quad p(\theta^{(p)}|\gamma^{(p)}) p(\theta^{(m)}|\gamma^{(m)}) p(\theta^{(d)}|\gamma^{(d)}) d\lambda d\theta^{(p)} d\theta^{(m)} d\theta^{(d)}
\end{aligned}
$$

We use a indicator function $I(\cdot) : Bool \to \{0, 1\}$. It returns 1 iff the statement is true. Then we can further expand the above term:

$$
\begin{aligned}
&= \int p(z|\lambda) p(\lambda|\alpha) d\lambda \times \prod_{r=1}^{R} \int\int\int \\
&\quad \prod_{n=1}^{N} p(S_n|\theta^{(p,r)}, \theta^{(m,r)}, \theta^{(d,r)})^{I(z_n=r)} \\
&\quad p(\theta^{(p)}|\gamma^{(p)}) p(\theta^{(m)}|\gamma^{(m)}) p(\theta^{(d)}|\gamma^{(d)}) d\lambda d\theta^{(p)} d\theta^{(m)} d\theta^{(d)}
\end{aligned}
$$

Define a count operation $C_{r,*,*,*}$ which represents the number of all the records in S whose phenotype index is r. We can further define $C_{r,i,*,*}$ to represents the number of all the records in S whose phenotype index is r and the patient id is i. Similarly we have $C_{r,*,j,*}$ and $C_{r,*,*,k}$. Then:

$$
\begin{aligned}
&= \int p(\lambda|\alpha) \prod_{r=1}^{R} \lambda_r^{C_{r,*,*,*}} d\lambda \\
&\quad \times \prod_{r=1}^{R} \int p(\theta^{(p,r)}|\gamma^{(p)}) \prod_{i=1}^{N_p} (\theta_i^{(p,r)})^{C_{r,i,*,*}} d\theta^{(p,r)} \\
&\quad \times \prod_{r=1}^{R} \int p(\theta^{(m,r)}|\gamma^{(m)}) \prod_{j=1}^{N_m} (\theta_i^{(m,r)})^{C_{r,*,j,*}} d\theta^{(m,r)} \\
&\quad \times \prod_{r=1}^{R} \int p(\theta^{(d,r)}|\gamma^{(d)}) \prod_{k=1}^{N_d} (\theta_i^{(d,r)})^{C_{r,*,*,k}} d\theta^{(d,r)}
\end{aligned}
$$

Plugin the Dirichlet prior, the first term can be further simplified as:

$$
\begin{aligned}
&\int \frac{\mathcal{T}(\alpha)}{\prod_{r=1}^{R} \mathcal{T}(\frac{\alpha}{R})} \prod_{r=1}^{R} \lambda_r^{\frac{\alpha}{R}-1} \prod_{r=1}^{R} \lambda_r^{C_{r,*,*,*}} d\lambda \\
=&\frac{\mathcal{T}(\alpha)}{\prod_{r=1}^{R} \mathcal{T}(\frac{\alpha}{R})} \int \prod_{r=1}^{R} \lambda_r^{C_{r,*,*,*}+\frac{\alpha}{R}-1} d\lambda \\
=&\frac{\mathcal{T}(\alpha)}{\prod_{r=1}^{R} \mathcal{T}(\frac{\alpha}{R})} \frac{\prod_{r=1}^{R} \mathcal{T}(C_{r,*,*,*} + \frac{\alpha}{R})}{\mathcal{T}(N+\alpha)} \\
&\int \frac{\mathcal{T}(N+\alpha)}{\prod_{r=1}^{R} \mathcal{T}(C_{r,*,*,*} + \frac{\alpha}{R})} \prod_{r=1}^{R} \lambda_r^{C_{r,*,*,*}+\frac{\alpha}{R}-1} d\lambda \\
=&\frac{\mathcal{T}(\alpha)}{\prod_{r=1}^{R} \mathcal{T}(\frac{\alpha}{R})} \frac{\prod_{r=1}^{R} \mathcal{T}(C_{r,*,*,*} + \frac{\alpha}{R})}{\mathcal{T}(N+\alpha)}
\end{aligned}
$$

$$
\begin{aligned}
\propto&\frac{\prod_{r=1}^{R} \mathcal{T}(C_{r,*,*,*} + \frac{\alpha}{R})}{\mathcal{T}(N+\alpha)} \\
=&\frac{\mathcal{T}(C_{z_t,*,*,*}^{-t} + \frac{\alpha}{R} + 1) \times \prod_{r \neq z_t} \mathcal{T}(C_{r,*,*,*} + \frac{\alpha}{R})}{\mathcal{T}(N+\alpha)} \\
=&(C_{z_t,*,*,*}^{-t} + \frac{\alpha}{R}) \frac{\mathcal{T}(C_{z_t,*,*,*}^{-n} + \frac{\alpha}{R}) \times \prod_{r \neq z_t} \mathcal{T}(C_{r,*,*,*} + \frac{\alpha}{R})}{\mathcal{T}(N+\alpha)} \\
\propto&C_{z_t,*,*,*}^{-t} + \frac{\alpha}{R}
\end{aligned}
$$

We use $\mathcal{T}(x+1) = x\mathcal{T}(x)$ here. $C_{z_t,*,*,*}^{-t}$ means the counting result without the t-th record. So $C_{z_t,*,*,*}^{-t} + 1$ equals to $C_{z_t,*,*,*}$. Similarly, we can simplify the second, third and fourth term, using the above tricks:

$$
\begin{aligned}
&\prod_{r=1}^{R} \int p(\theta^{(p,r)}|\gamma^{(p)}) \prod_{i=1}^{N_p} (\theta_i^{(p,r)})^{C_{r,i,*,*}} d\theta^{(p,r)} \\
\propto&\frac{C_{z_t,p_t,*,*}^{-t} + \gamma_{p_t}^{(p)}}{\sum_{i=1}^{N_p} C_{z_t,i,*,*}^{-t} + \gamma_i^{(p)}}, \\
&\prod_{r=1}^{R} \int p(\theta^{(m,r)}|\gamma^{(m)}) \prod_{j=1}^{N_m} (\theta_i^{(m,r)})^{C_{r,*,j,*}} d\theta^{(m,r)} \\
\propto&\frac{C_{z_t,*,m_t,*}^{-t} + \gamma_{m_t}^{(m)}}{\sum_{j=1}^{N_m} C_{z_t,*,j,*}^{-t} + \gamma_j^{(m)}} \text{ and} \\
&\prod_{r=1}^{R} \int p(\theta^{(d,r)}|\gamma^{(d)}) \prod_{k=1}^{N_d} (\theta_i^{(d,r)})^{C_{r,*,*,k}} d\theta^{(d,r)} \\
\propto&\frac{C_{z_t,*,*,d_t}^{-t} + \gamma_{d_t}^{(d)}}{\sum_{k=1}^{N_d} C_{z_t,*,*,k}^{-t} + \gamma_k^{(d)}}
\end{aligned}
$$

So

$$
\begin{aligned}
&p(z_t|z_{-t}, S, \alpha, \gamma^{(p)}, \gamma^{(m)}, \gamma^{(d)}) \\
\propto&(C_{z_t,*,*,*}^{-t} + \frac{\alpha}{R}) \times \frac{C_{z_t,p_t,*,*}^{-t} + \gamma_{p_t}^{(p)}}{\sum_{i=1}^{N_p} C_{z_t,i,*,*}^{-t} + \gamma_i^{(p)}} \\
&\times \frac{C_{z_t,*,m_t,*}^{-t} + \gamma_{m_t}^{(m)}}{\sum_{j=1}^{N_m} C_{z_t,*,j,*}^{-t} + \gamma_j^{(m)}} \times \frac{C_{z_t,*,*,d_t}^{-t} + \gamma_{d_t}^{(d)}}{\sum_{k=1}^{N_d} C_{z_t,*,*,k}^{-t} + \gamma_k^{(d)}}
\end{aligned}
$$

Since $\sum_{r=1}^{\infty} C_{z_t,*,*,*}^{-t} + \frac{\alpha}{R} = N + \alpha - 1$, when $R \to \infty$, the weight of existing mixture is $\frac{C_{z_t,*,*,*}^{-t}}{N+\alpha-1}$, and the weight of new mixture is $\frac{\alpha}{N+\alpha-1}$.

III. Algorithm

A. Single thread Gibbs sampling

Now since we have the marginal posterior probability, we can run Gibbs sampling to inference the hidden variables $z_n, n = 1, \ldots, N$. Define $f(t) = \frac{C_{z_t,p_t,*,*}^{-t} + \gamma_{p_t}^{(p)}}{\sum_{i=1}^{N_p} C_{z_t,i,*,*}^{-t} + \gamma_i^{(p)}} \times \frac{C_{z_t,*,m_t,*}^{-t} + \gamma_{m_t}^{(m)}}{\sum_{j=1}^{N_m} C_{z_t,*,j,*}^{-t} + \gamma_j^{(m)}} \times \frac{C_{z_t,*,*,d_t}^{-t} + \gamma_{d_t}^{(d)}}{\sum_{k=1}^{N_d} C_{z_t,*,*,k}^{-t} + \gamma_k^{(d)}}$. The algorithm is shown below:

1. Set $R = R_0$, where R_0 is the initial number of phenotypes; Randomly initialize z.

2. Count the values $C_{(\cdot,\cdot,\cdot,\cdot)}$.
3. For $t \in 1, 2, \ldots, N$
   a) Remove $S_t = (p_t, m_t, d_t)$ from data. Update $C_{(\cdot,\cdot,\cdot,\cdot)}$
   b) If S_t is the only instance in mixture z_t, then remove this mixture, and reduce R by one.
   c) Draw z_t:

$$p(z_t = r \leq R) \propto \frac{C^{-t}_{z_t,*,*,*}}{N + \alpha - 1} \times f(t)$$
$$p(z_t = R + 1) \propto \frac{\alpha}{N + \alpha - 1} \times f(t)$$

4. Update R if $z_t = R + 1$.
5. Update $C_{(\cdot,\cdot,\cdot,\cdot)}$ accordingly.

IV. Experiments

V. Conclusion

Acknowledgment

References

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.